U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Condition Monitoring of Bearings under Low and Medium Speed Rotation

Paulo Jorge Nascimento Morais

Porto, July 2016

Faculdade de Engenharia da Universidade do Porto
Departamento de Engenharia Mecânica e Gestão Industrial

Condition Monitoring of Bearings under Low and Medium Speed Rotation

Paulo Jorge Nascimento Morais

Dissertation developed at Siemens SISW in Leuven and submitted to Faculdade de Engenharia da Universidade do Porto for the degree of Master in Mechanical Engineering

Supervisor at FEUP:
Prof. José Rodrigues

Supervisor at Siemens:
Eng. Carina Freitas

Porto, July 2016

LaTeX

Condition Monitoring of Bearings under Low and Medium Speed Rotation
P. J. Morais
2016
FEUP-U.PORTO

To my family

Acknowledgements

The work that will be presented here is the combination of knowledge and experience obtained with the help of several individuals, who were part of my journey leaving their valued contribution, and to whom I would like to express my acknowledgement.

First of all, to Faculdade de Engenharia da Universidade do Porto, not only it had been my second home during the past 5 years but also a place where I grown personal and professionally. I would like to adress all the professors and staff that everyday contribute to keep FEUP as one of the most prestigious and respected schools in the country. To professor José Dias Rodrigues, who was my internal supervisor and awaken my curiosity to the field of mechanical vibrations. Also to professor José Pedro Reina, who since my beginning in this course was always available to help and whose perfectionism inspired me.

To Siemens SISW, in the person of the Eng. Carina Freitas, who provided the means for this research and that welcome me as one of their own, allowing me to experience the real life work. To all my colleagues in the company, who made everyday a continuous learning opportunity.

To all my friends, Inês Baptista, Hélder Araújo, Francisco Miranda, Tiago Fraga and Ricardo Salvado, for all the support and the good times that they provided in this journey. To Miguel Palmares, for the advise, the patience and the unconditional friendship. To Bruno Vieira, the first person I met in Porto, for his brotherhood and partnership in every challenge never forgetting the life advice when needed. To Catarina Costa, for the patience, the support, the love and the courage that she gave to me. To all my friends that one way or another, were present in my life and that made me a better person.

To my parents, my sister and my grandparents, for all the unconditional support, for all the words, the affection, the comprehension and the love that they gave me. I am eternally grateful to them and could ever return everything they do for me, not only in the last five years, but in my whole life, taking care of me and making me better every day.

To all of those who could not see me finishing this goal but that made me laugh as there was anything else.

Abstract

Rolling element bearings are present in the most part of the mechanical equipments both domestic and industrial. Their failure is responsible for the majority of the downtimes and consequently loss of production and profit. An appropriate condition monitoring technique minimizes the catastrophic failure of important equipment allowing an effective scheduling of the maintenance actions. Vibration monitoring is one of the most used techniques on condition monitoring. Despite other techniques, such as wear debris or termography, return also good information, vibration monitoring strategies are widely spread on condition monitoring due primarily to the accuracy of the measure information that enables the detection of faults in an incipient stage, but also to the facility of measuring, collecting and processing data. High rotational speeds are a well studied topic with different methods, such as the envelope analysis or the wavelet transform, performing fault detection in an easy and efficient way. However, for the low and medium speed regimes, where the wind turbines work for example, the same is not true and there are serious difficulties which need to be surpassed. For this regimes, the slow evolution in time and the weak amplitude of the induced fault vibrations are the major difficulties. This text presents a review on the condition monitoring basis, the most used processing techniques to improve the fault detection and an approach on the trending of time signal features with the speed. The goal is to evaluate different features and conclude which are the ones that return better results. Some results are presented regarding a low and a medium speed rotational case where an evaluation of the envelope analysis and the features extraction was conducted.

Resumo

Os rolamentos de esferas estão presentes em grande parte dos equipamentos mecânicos tanto domésticos como industriais. A falha destes componentes é responsável pela maioria das quebras de produção e consquentemente de lucro. Uma estratégia de monitorização adequada minimiza as falhas catastróficas de equipamentos importantes e permite um planeamento adequeado das acções de manutenção. A monitorização de vibrações é uma das mais usadas técnicas no acompanhamento da condição dos equipamentos. Apesar de outras técnicas tais como a análise de detritos ou a termografia, permitirem também uma monitorização eficaz, a análise de vibrações é a mais amplamente utilizada devido principalmente à qualidade da informação obtida, que permite a deteção de falhas numa fase incial, mas também pela facilidade em realizar medições, colectar e processar os resultados. As elevadas velocidades de rotação são um tópico dominado por vários métodos, tais como a análise do envelope ou a transformada de wavelet, que detectam fácil e eficazmente a presença de uma falha. Contudo, no regime das baixas e médias velocidades, como é o caso das turbinas eólicas, o mesmo não se pode afirmar e existem inúmeras dificuldades que precisam de ser ultrapassadas. Nestes casos, a lenta evolução com o tempo e a fraca amplitude das vibrações geradas pela falha representam o maior problema. Este texto apresenta uma revisão dos aspectos fundamentais da monitorização da condição de equipamentos, as mais usadas técnias de processamento que melhoram a probabilidade de deteção da falha e uma análise da tendência de parâmetros estatísticos do sinal com a velocidade. O objectivo é avaliar diferentes parâmetros e perceber quais os que apresentam melhores resultados. No final são apresentados alguns resultados relativos a baixas e médias velocidade de rotação analisando o envelope e os parâmetros extraídos do sinal.

Keywords

- Condition Monitoring
- Rolling Element Bearings
- Rotating Machinery
- Low Speed Rotation
- Medium Speed Rotation
- Vibration Monitoring
- Envelope Analysis
- Features Extraction
- Minimum Entropy Deconvolution

Contents

List of Figures

List of Tables

List of Symbols

α	misalignment between shafts	rad
ΔdB	Difference dB-Spectra filter	/
Δt	sampling interval	s
η	viscosity	Pa s
λ	lubrication regime indicator	/
λ_i	i-th eigenvalue of the covariance matrix	/
$\mathcal{F}$	discrete Fourier transform	/
$\mathcal{H}$	Hilbert transform	/
$\mathcal{M}(f)$	Matched filter	/
$\mathcal{W}(f)$	Wiener filter	/
$\mathcal{W}(t)$	window moved along the signal	/
$\mathscr{W}_\Psi$	continuous wavelet transformation	/
μ	friction coefficient	/
ω	angular velocity	$\mathrm{rad\,s^{-1}}$
ω_C	angular velocity at each ball center	$\mathrm{rad\,s^{-1}}$
ω_I	angular velocity of the inner race	$\mathrm{rad\,s^{-1}}$
ω_n	natural frequency	$\mathrm{rad\,s^{-1}}$
ω_O	angular velocity of the outer race	$\mathrm{rad\,s^{-1}}$
$\omega_{1,2}$	input and output shaft angular velocity	$\mathrm{rad\,s^{-1}}$
ϕ	pressure angle	rad
$\Psi(t)$	mother wavelet	/
$\rho(A,B)$	correlation between two variables	/
θ	input shaft rotation angle	rad
υ	velocity	$\mathrm{m\,s^{-1}}$
$\Psi(a,b)$	wavelet transform	/
ξ	damping coefficient	/
a_n, b_n	Fourier series coefficients	/
cov	covariance between two variables	/
D	pitch diameter	m
d	ball diameter	m
e	distance between the mass and the rotation center	m
f_S	shaft speed	Hz
h	film thickness	m
M	system mass	kg
m_0	unbalanced mass	kg
N	number of rolling elements	/
N_C	rotational speed of the carrier	rpm
N_G	rotational speed of the gear	rpm
N_P	rotational speed of the pinion	rpm
N_R	rotational speed of the outer ring	rpm
N_S	rotational speed of the sun	rpm
P	load on the contact	N
$P(x)$	probability distribution function	/
$p(x)$	probability density	/
p_i	i-th eigenvector of the covariance matrix	/
R	correlation matrix of two variables	/
r	ratio between angular velocity and natural frequency	/

R_A	combined surface roughness	m
r_B	ball radius	m
r_I	inner race radius	m
r_R	semi-width of the rough contact	m
r_S	semi-width of the ideal contact	m
R_{DIN}	roughness indicator	m
$S(f,\tau)$	short time Fourier transform	/
T	torque on the input shaft	N m
T_0	time between pulses	s
T_G	number of gear tooth	/
T_P	number of pinion tooth	/
T_R	number of outer ring tooth	/
T_S	number of sun tooth	/
V_B	velocity of a point on the ball surface	$\mathrm{m\,s^{-1}}$
V_C	velocity at each ball center	$\mathrm{m\,s^{-1}}$
V_I	velocity of the inner race	$\mathrm{m\,s^{-1}}$
V_O	velocity of the outer race	$\mathrm{m\,s^{-1}}$

List of Acronyms

CBMS Condition Based Monitoring Strategy 1

AE Acoustic Emission 3

CM Condition Monitoring 5

VA Vibration Analysis 5

ISO International Organization for Standardization 6

DOF Degree of Freedom 13

API American Petroleum Institute 14

TE Transmission Error 11

FPF Fundamental Periodic Frequency 19

RMS Root Mean Square 20

FCF Fault Characteristic Frequencies 9

BPFO Ball Pass Outer Frequency 9

BPFI Ball Pass Inner Frequency 9

BSF Ball Spin Frequency 10

FTF Fundamental Train Frequency 10

DIN Deutsches Institut für Normung 26

EWEA European Wind Energy Association 40

WWEA World Wind Energy Association 40

SCADA Supervisory Control and Data Acquisition 41

MEMS Micro Electro-Mechanical Systems 31

AC Alternated Current 31

DC Direct Current 31

IEPE Integrated Electronic Piezoelectric 31

CF Crest Factor 84

FFT Fast Fourier Transform 49

PP Peak-to-Peak 83

Chapter 1

Introduction

1.1 Condition Monitoring

The mechanical equipments sooner or later develop faults on their components that will led to failure and the need of repairs. Since the beginning of mechanical engineering, monitoring is conducted through all the equipments even if unconsciously, every time a maintenance responsible passes through a machine and an experienced one could tell if there are problems with it only by listening the machine noise. Different types of strategies may be adopted to perform the maintenance of a industrial plant and a long way since the early strategies allow modern and reliable techniques to improve the safety and profit of the equipments.

A wide range of techniques and fields are involved when it comes to condition monitoring. The methods vary from chemical effects, to wear debris, passing through termography, performance analysis or vibration analysis. The last one is the most spread and commonly used technique adopted in the maintenance strategies. Even when another technique is adopted, vibration monitoring is somehow part of the strategy as an additional information when it comes to make decisions. The general way to perform vibration monitoring is by envelope analysis which consists of analysing the envelope spectrum of the demodulated signal. However, techniques such as cepstrum editing or minimum entropy deconvolution were introduced to improve the final results.

Adopting a Condition Based Maintenance Strategy

A strategy of condition monitoring, also called predictive maintenance is based on being able to assess the equipment state and predict future failures during running [1]. This approach allows the user to make decisions which will reduce the downtimes by avoiding unexpected catastrophic failures and also to schedule maintenance actions to be taken at the optimum time according to the needs of the machine.

> *"Condition Monitoring is a monitoring process or a sensitive tool that focus on early detection of faults, failures and wear of machinery with the intention to minimize downtimes and maintenance costs, and consequently, maximize production"*[2]

Despite all the advantages that a Condition Based Monitoring Strategy (CBMS) has, it still requires a carefully processing and results interpretation combined with the analyst expertise. Due to this, there are currently different algorithms and techniques for data processing which provide different levels of accuracy in fault detection [3]. Condition monitoring has been implemented through the past 30-40 years and despite its initial limitations, over the past 15 years the results provided by this approach are mainly recognized as the best maintenance

strategy in most cases, such as hydroelectric power plants and wind turbines[1]. Before condition monitoring was introduced in the maintenance strategies, engineers used to adopt other methods that will now be discussed superficially. The *preventive maintenance* is established at regular intervals that have to be shorter than the time between failures and usually assure that 98-99% of the equipments will not experience failure.

The drawbacks of this method are that some components could have run for more time while others are replaced at a tardy stage which can led to failures that otherwise would never occurred. This suits applications where the components tend to wear at a constant and predictable rate, that can be calculated if the operational regime is steady[1]. On the other hand, some components such as rolling bearings present some spread on their wearing and therefore a predictive maintenance strategy must not be applied, however this topic will be further discussed in this text. The most natural maintenance strategy would be the $run - to - break$ and as one can infer it consist of run the machine until it breaks [1]. This solution is used in industries where the losses for breakdown are not high and also when the machine is not essential for production. One example of this is the sewing machines in the textile industry. Despite providing the larger time between downtimes, this technique also allows catastrophic failures which can damaged other components or machines in the proximity.

Economic Benefits

Nowadays, the demanding for products and solutions which are more efficient, cheaper and reliable is increasing as society evolves and therefore the consumption habits. In order to minimize the costs, a CBMS should allow fault detection, diagnosis and prognosis for example to predict the remaining useful time of an equipment. An effective maintenance strategy allows companies to reduce downtime and also schedule more accurate interventions on their equipments. Table 1.1 indicates the major causes of accident occurred between 1958 and 1987 in the petrochemical industry by poor maintenance practices.

Table 1.1: Causes for machine failure [4]

Causes	Frequency (%)
Mechanical failure	38
Operation error	26
Unknown	12
Process upset	10
Natural hazards	7
Design error	4
Sabotage	3

In the past, adopting a maintenance strategy was seen as overhead and for some companies the solution would achieve its objective if it remained within its budget. According to Davies [4], in many companies, maintenance is still viewed as a technical activity and monitoring strategies are not fully recognized. A study conducted in 1988 by the Department of Trade and Industry in the United Kingdom showed that Britain's factories could save around 16 billion € per year with proper strategies applied. In accordance with Davies, *RollsRoyce* observed that 70% of their engine failures were due to debris contamination. In the 90's the costs of failures were estimated around 50% of the quality costs. Apart from the components cost, a machine running in a poor condition also magnify the energy bill and at the same time it becomes less environmentally friendly. *Texaco's Pembroke* saved around 600 000 € per year

by adopting an effective energy monitoring and management program [4].

Other Maintenance Strategies

This text uses vibration analysis as the principal way to detect faults. The basis is that a machine produces a certain signature that will change in the presence of a defect. However, other approaches can be used, such as lubricant analysis, termography, Acoustic Emission (AE), ultrasound, electrostatic or shock-pulse [5].

Besides the vibration signature, important information can also be extracted from the lubricant. The goal is to analyse the presence of debris or external contaminants. A magnet or *chip detector*, can be mounted on the system to retain metallic particles and $ferrography$ or *spectographic* analysis can be conducted on lubricant samples [1]. Combining these techniques allows the analysis, in terms of quantity, type, shape and size of particles, being them metallic or non-metallic. Termography is used to monitor $quasi - static$ situations for instance in containers for hot fluids. This technique uses the increase of temperature as a parameter to detect faults. Regardless its applicability in some situations, its not suitable for bearings, because the rise in temperature for them happens in the last stages of life where the fault detection is to late[1].

1.2 Motivations

Rolling element bearings, as one of the most used components in machinery represent a crucial part of the maintenance strategies and are the main focus of this thesis. However, despite the facility in applying the vibration monitoring to bearings running at high rotational speeds, the same is not true for the low and medium regimes where some serious difficulties need to be surpassed. For this regimes, the slow evolution in time and the weak amplitude of the induced fault vibrations are the major difficulties. A typical application of what will be presented in this text are the wind turbines which run at very low speed, the blades run at around 20 RPM which leads the bearings where their shaft is supported to be of difficult analysis. In fact, these components are responsible for up to 70% of the gearbox failures as will be seen.

A rapid way to assess the bearing condition would be by analysing simple parameters such as the root mean square, the variance, the kurtosis or the skewness. Those and other features are presented below and other authors tried to combine several of them to create new condition parameters. Despite some organizations such as the International Standard Organization propose alert levels based on the level of some specific statistical parameters, for low speed rotation there is not a common opinion, with the alert values being fixed as a constant for all the regime. With this in mind, some objectives were defined for this thesis which are enumerated as follows.

1.3 Objectives

- Understand the fundamentals of condition monitoring, through different techniques, and the state of the technology. Clarify the use of vibration analysis in this area and where it fits with other techniques.

- Review the knowledge on bearing diagnosis, specifically for low and medium operational speeds, the fault propagation and the aspects which influence it. Understand the procedures to perform correct measurements and to minimize the external interferences.
- Assess the applicability of the condition monitoring strategies in industrial applications namely for rolling element bearings.
- Study the processing techniques available for vibration monitoring and clarify their advantages when it comes to bearings. Apply the envelope analysis to real cases and discuss the final results.
- Collect different features to be extracted from the time signal and evaluate their evolution with the rotational speed and the load applied.
- Find relations between the features which can be used as an information to create a condition parameter as a combination of different features.

1.4 Structure

This text is structured in seven chapters. The next two clear the fundamental aspects on the condition monitoring, namely on vibrations analysis and rotating machines, the diagnosis of bearings under low-speed rotation and the common defects and interferences. The fourth chapter detail the data processing techniques available for the condition analyst and effective procedures to clear and enhance the signal namely the cepstrum editing and the minimum entropy deconvolution. After that, in the fifth chapter, are presented the features studied and the correlation conducted between them. The chapter six present the results obtained for the two cases studied of bearings under low and medium speed rotations, processed with both envelope analysis and features extraction. For last but not least the conclusions of this work, future needs in the field and opportunities.

Chapter 2

Vibration Fundamentals

2.1 Vibration Monitoring

Mechanical systems are present in almost all the factories and because of that, vibration monitoring is generally used as a critical part of a Condition Monitoring (CM) strategy. Vibration Analysis (VA) is the most common technique used to monitor rotating equipments therefore a powerful tool for bearing fault diagnosis as we will see further in this text [2]. Vibrations are generated every time a machine is running, however CM allows the user to separate faulty vibrations from the ones which are typical from the equipment or the machine signature. Some of those vibrations are connected to periodic events arising from components such as shafts, bearings, gears or rotating electric fields [1].

Through vibration signals, one is able to process the information on different domains such as time, frequency or time-frequency domain to detect different fault types and locations. The periodic events masked on the time domain signal can be easily detected once the signal is represented on the frequency domain and usually allows a direct access to information of the fault source. Hence many algorithms use this representation as a base for their diagnostic. On the other hand, there are other phenomena which are not completely periodic, for instance the vibrations produced in internal combustions engines where the number of combustion events occur at each engine cycle [1]. Nonetheless, this is an advantage as it is possible to separate the non-periodic signals from the periodic ones. For varying operational conditions the analysis of the vibration signal becomes harder because of the load and speed influence on the measured signal [6]. CM strategies based on vibration signature analysis premise two facts [4]:

- Every failure mode in study have distinct vibration frequency components that can be isolated and identified.

- The amplitude of each component signature will remain constant unless a change occurs in its operating dynamics.

Some parameters, such as sensor characteristics, pre and post-processing will affect the final diagnosis. Between the vibration source and the electrical signal acquired by the user, there are factors which influence the information. One clear difference is between the absolute vibration of a machine housing and the relative vibration of the bearing and the housing. In the last case, there is a film thickness separating the shaft and the housing [1].

Benefits of Vibration Analysis

Currently there are a wide variety of International Organization for Standardization (ISO) standards which establish acceptable vibration limits. These standards specify methodologies for measurement, handling and processing data to perform machines diagnosis and prognosis. Two of those standards are the ISO/7919 *Mechanical vibration of non-reciprocating machines - Measurements on rotating shafts and evaluation criteria* and the ISO/10816 *Mechanical vibration - Evaluation of machine vibration by measurements on non-rotating parts* [5]. The advantages of the VA compared with other methods is the fact that it reacts immediately to changes in the machine behaviour and can be used as a permanent or intermittent technique. If compared with oil analysis, VA provide better results on the fault location and detection by the fact that the lubricant samples have to be processed in specific equipment and the vibration signals can be processed on site. In addition to this, the tools available nowadays allow the very weak fault impulses to be detected even in the presence of strong noise and interference [1].

Some equipments have permanent transducers in order to be continuously monitored. The advantages of this technique is the ability to react instantaneously to changes which allow better protecting against catastrophic failures, minimizing the costs. In situations where the faults cannot be predicted this is the best solutions to adopt. An example gave by [1] is the sudden unbalance that can occur on fans handling dirty gas, where there is normally a deposit of material on the blades over time. The distribution is usually uniform, but if for some reason some deposits detach from the surface this can led to a sudden unbalance. On the other hand, the investment in a permanent monitoring system is only suitable for critical machines. In opposition to a permanent monitoring system, using an intermittent monitoring approach allows to reduce the cost of the equipment and should be applied primarily when the cost of lost production from downtimes, outweighs the cost of the machine itself. Should not be used in cases where unpredictable failures may occur [1]. There has to be a balance between the frequency of monitoring and the time between failures to assure the minimum losses. This type of technique is used on a vaster way than the permanent monitoring.

2.2 Rotating Machines

In order to proper analyse the vibration signals we need to eliminate other interferences that could be present in the machine. Some faults are observed at the speed of the shaft and its harmonics or sub harmonics. The typical cases are unbalance, misalignment and bent which will be discussed in this section.

2.2.1 Rolling Element Bearings

Rolling element bearings are found widely distributed in both domestic and industrial applications such as turbines, compressors, pumps or motors. These components are critical for the machine operation and their main function is to support a rotor system. A bearing, transfer the forces from rotating parts to the main structure and is designed to provide less friction at the supports. The two most common types of bearings used are the anti-friction bearings or rolling element bearings and the journal bearings. The latter, works on the principle of hydrodynamic lubrication. These are filled with a lubricating oil, characterized by a certain viscosity and while running at high speeds, a pressure is built up toward the centre of the shaft. This fluid pressure, is responsible for supporting the load on the shaft and helps reducing the friction and consequently the wear. To achieve the hydrodynamic lubrication regime, the bearing has

to pass through the mixed and boundary lubrication regimes. While starting and stopping the material used in these bearings is subject to severe degradation and a solution called hydrostatic bearings minimize the problem. These bearings are provided with external pressurized lubricant and are used when the loads are very high. Journal bearings, have a damping ratio provided by the lubricant, therefore the vibration response at any speed can be controlled [7].

The other type of bearings, the rolling element, are the most widely used ones and will be the studied in this text. Generally, bearings are the most precise machine parts and are typically made with tolerances that are around ten times more strict than the other machine components [8]. Despite this fact, only about 10 to 20% of the bearings reach their design life due to external influencing factors. According to [9], bearing faults account for up to 40% of motor faults. Among others, some of the external factors that reduce the bearing life are improper lubricant, foreign particles, moisture intrusion, excessive vibration and high dynamic loads or false brielling, during transport or when the machine is out of operation for a long time.

Rolling element bearings have four components, two circular races, designated inner and outer, between which the rolling elements are held. The rolling elements can have different geometries such as rollers, tapered rollers or spherical balls. The fourth component is the cage, which ensures that two rolling elements do not come into contact with each other. Figures 2.1 presents a scheme of the components of a rolling elements bearing [8]. The sealing on both ends of the component helps to minimize the amount of external particles that go into the track. The rolling elements are made of hardened steel and are manufactured as it had been said with very strict tolerances and very high surface finish. However due to manufacturing imperfections these components are non-circular and some type of radial variations are always present even if only a few microns. Due to this, when the bearing is in motion, a vibration wave is generated in it. Currently, the best bearings manufactured in the world have radial imperfection between 3 to 5 μm according to [7]. Due to this fact, the radial load on the bearing is constantly changing with the rotation of the shaft, what makes the bearing vibration signal amplitude modulated. To perform bearing diagnosis, the signal has therefore to be demodulated and analysed in the frequency domain as is further detailed in Chapter 4.

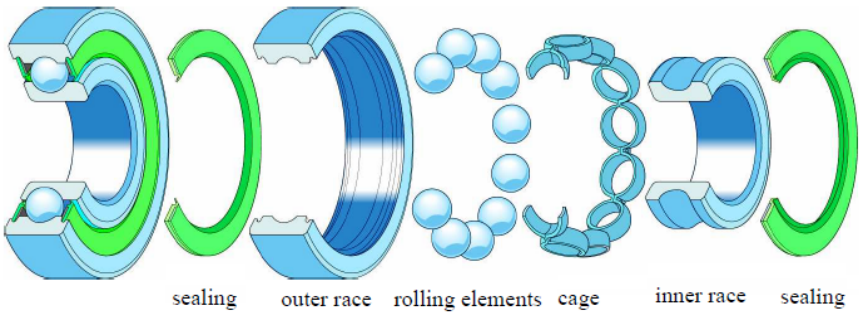


Figure 2.1: Rolling element bearing components [8]

Figure 2.2 shows how the faulty signals are generated by localized faults in the various components of a rolling element bearing and the correspondent envelope signal by amplitude demodulation. When the rolling element strikes a local fault, a shock is introduced, exciting high-frequency resonances of the whole structure between the origin and the transducer. The broadband excited by the shocks is modulated in amplitude by two factors according to Randall [10].

1. The strength of the bursts depend on the load applied to the rolling elements and is normally modulated by the rate at which the fault is passing through the load zone.

2. For moving faults, the transfer function of the transmission path varies with respect to the fixed positions of the response transducers.

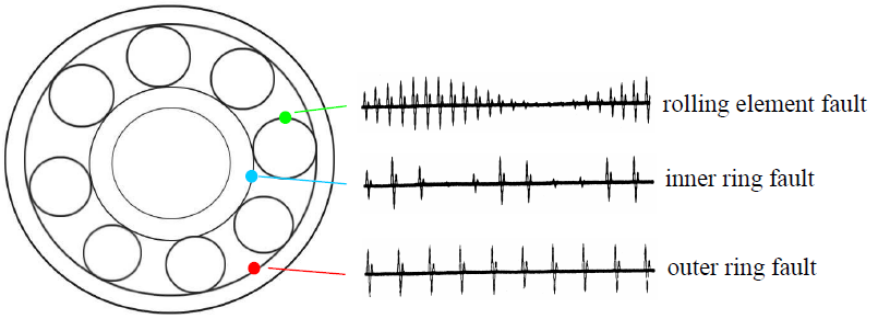


Figure 2.2: Signal generation on a faulty bearing [8]

2.2.1.1 Fault Characteristic Frequencies

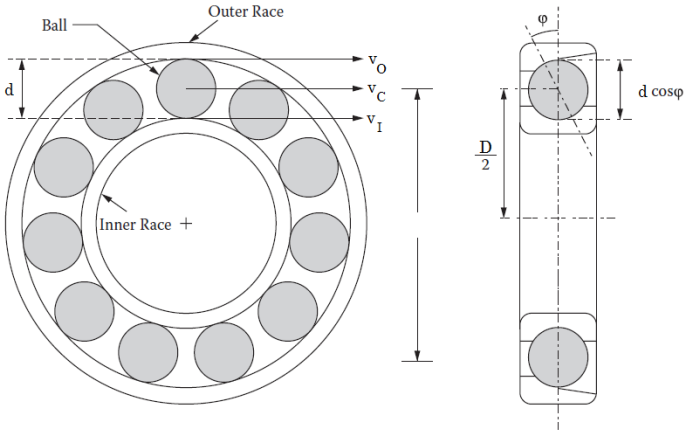


Figure 2.3: Geometry of a rolling element bearing [7]

The characteristic bearing defect frequencies for different rolling elements can be determined and their deduction will now be performed. Figure 2.3 helps the reader to identify the different parameters used in the following equations. Designating the velocity at each ball centre, the inner race and outer race velocities by V_C, V_I and V_O respectively, the ball diameter by d and the pitch diameter by D, the velocity V_C is given by equation (2.1).

$$V_C = \frac{V_I + V_O}{2} \tag{2.1}$$

The velocities V_I and V_O can be written in terms of the angular speed of the inner ω_I and outer race ω_O.

$$V_I = \omega_I \left(1 - \frac{d\cos\phi}{D}\right) \tag{2.2}$$

$$V_I = \omega_O \left(1 + \frac{d\cos\phi}{D}\right) \tag{2.3}$$

Where ϕ represents the pressure angle where the load is being applied. Equation (2.1) can now be rewritten according to equation (2.2) and (2.3).

$$\omega_C = \frac{1}{2}\left[\omega_I \left(1 - \frac{d\cos\phi}{D}\right) + \omega_O \left(1 + \frac{d\cos\phi}{D}\right)\right] \tag{2.4}$$

The Ball Pass Outer Frequency (BPFO), that is the frequency at which a ball passes a fixed point on the outer race, is given by the equation (2.5) where N represents the number of rolling elements in the bearing.

$$BPFO = N|\omega_C - \omega_O| \tag{2.5}$$

Substituting equation (2.4) into (2.5).

$$BPFO = N \left| \frac{1}{2}\left[\omega_I \left(1 - \frac{d\cos\phi}{D}\right) + \omega_O \left(1 + \frac{d\cos\phi}{D}\right)\right] - \omega_O \right| \tag{2.6}$$

$$= \frac{N}{2} \left| \omega_I \left(1 - \frac{d\cos\phi}{D}\right) + \omega_O \left(1 + \frac{d\cos\phi}{D}\right) - 2\cdot\omega_O \right| \tag{2.7}$$

$$= \frac{N}{2} \left| \omega_I \left(1 - \frac{d\cos\phi}{D}\right) - \omega_O \left(1 - \frac{d\cos\phi}{D}\right) \right| \tag{2.8}$$

$$= \frac{N}{2} \left| (\omega_I - \omega_O) \left(1 - \frac{d\cos\phi}{D}\right) \right| \tag{2.9}$$

Applying the same procedure for the Ball Pass Inner Frequency (BPFI), that is the frequency at which a ball passes a fixed point in the inner race and is given by equation (2.10).

$$BPFI = N|\omega_I - \omega_C| \tag{2.10}$$

Similarly to what was done for the BPFO, it is possible to use the equations (2.4) to (2.10) and get the Fault Characteristic Frequencies (FCF) as functions of ω_I and ω_O.

$$BPFI = N \left| \omega_I - \frac{1}{2}\left[\omega_I \left(1 - \frac{d\cos\phi}{D}\right) + \omega_O \left(1 + \frac{d\cos\phi}{D}\right)\right] \right| \tag{2.11}$$

$$= \frac{N}{2} \left| 2\cdot\omega_I - \omega_I \left(1 - \frac{d\cos\phi}{D}\right) - \omega_O \left(1 + \frac{d\cos\phi}{D}\right) \right| \tag{2.12}$$

$$= \frac{N}{2} \left| \omega_I \left(1 + \frac{d\cos\phi}{D}\right) - \omega_O \left(1 + \frac{d\cos\phi}{D}\right) \right| \tag{2.13}$$

$$= \frac{N}{2} \left| (\omega_I - \omega_O) \left(1 + \frac{d\cos\phi}{D}\right) \right| \tag{2.14}$$

Lets now focus on a point, in the inner race, in contact with the rolling element surface. Considering the rolling element as a ball and the linear velocity of a point on the inner race, the linear velocity V_B of a point on the ball surface is given by equation (2.15).

$$V_B = \left(\omega_I - \omega_C \right) r_I \tag{2.15}$$

Therefore, because $V_B = \omega \cdot r_B$ the Ball Spin Frequency (BSF) is given by the equation (2.17), where r_I and r_B represent the inner and ball radius respectively.

$$r_I = \frac{D - d\cos\phi}{2} \qquad\qquad r_B = \frac{d}{2} \tag{2.16}$$

$$BSF = \left| (\omega_I - \omega_C) \, \frac{D - d\cos\phi}{d} \right| \tag{2.17}$$

The value of $\omega_I - \omega_C$ can be written as follows.

$$\left(\omega_I - \omega_C \right) = \frac{1}{2} \left(\omega_I - \omega_O \right) \left(1 + \frac{d\cos\phi}{D} \right) \tag{2.18}$$

Using now the equations (2.4) and (2.18), the equation (2.17) can be rewritten.

$$BSF = \left| \frac{1}{2} \left(\omega_I - \omega_O \right) \left(1 + \frac{d\cos\phi}{D} \right) \cdot \frac{D - d\cos\phi}{d} \right| \tag{2.19}$$

$$= \frac{1}{2d} \left| \left(\omega_I - \omega_O \right) \left(1 + \frac{d\cos\phi}{D} \right) (D - d\cos\phi) \right| \tag{2.20}$$

$$= \frac{1}{2d} \left| \left(\omega_I - \omega_O \right) \left(D + \cancel{d\cos\phi} - \cancel{d\cos\phi} - \frac{(d\cos\phi)^2}{D} \right) \right| \tag{2.21}$$

$$= \frac{D}{2d} \left| (\omega_I - \omega_O) \left[1 - \left(\frac{d\cos\phi}{D} \right)^2 \right] \right| \tag{2.22}$$

$$\tag{2.23}$$

And finally, the Fundamental Train Frequency (FTF), given by the velocity equation of each ball centre

$$FTF = \frac{1}{2} \left[\omega_I \left(1 - \frac{d\cos\phi}{D} \right) + \omega_O \left(1 + \frac{d\cos\phi}{D} \right) \right] \tag{2.24}$$

In common applications, the outer race of the bearings is fixed and equations (2.9), (2.14), (2.22) and (2.24) for $\omega_O = 0$ are given by equations (2.25) to (2.28), where f_S represents the angular speed of the shaft in Hz.

$$BPFO = \frac{f_S \cdot N}{2} \left(1 - \frac{d\cos\phi}{D}\right) \tag{2.25}$$

$$BPFI = \frac{f_S \cdot N}{2} \left(1 + \frac{d\cos\phi}{D}\right) \tag{2.26}$$

$$BSF = \frac{f_S \cdot D}{2d} \left[1 - \left(\frac{d\cos\phi}{D}\right)^2\right] \tag{2.27}$$

$$FTF = \frac{f_S}{2} \left(1 - \frac{d\cos\phi}{D}\right) \tag{2.28}$$

These frequencies are the theoretical kinematic ones, assuming no slip. The reality is that there must always exist some slip due to the angle ϕ variation with the position of each rolling element in the bearing, as the ratio of local radial to axial load changes. Therefore, each rolling diameter is different which means that each ball is trying to roll at a different speed. The mean speed of all rolling elements is ensured by the cage, which introduces some random slip, varying between 1-2% from the calculated value and has random variation around the mean frequency [10]. Because of this random slip, the character of the signal changes, thus in order to properly detect faults, envelope analysis must be performed. The bearings, have a natural frequency due to their mass and stiffness and are usually designed so that their natural frequency is beyond the operating frequencies of the machine to avoid resonances. However, incipient faults, due to pits in the races, give rise to impact excitation of the races when the rolling elements go over those pits. These forces, which have short time duration are responsible for high frequency excitation that consequently induce resonant vibration in the races, producing a high level of vibration at high frequencies. Usually, these vibrations are over 6 kHz and help the detection of faults, in the early stages, by the presence of high frequency vibrations. Resonant AE sensors allow to measure in a certain frequency band located around the resonance frequency of the sensor. The sensor used for the results presented in the Chapter 6 had a demodulation function built in, however this is not a common practice in real life applications.

2.2.2 Vibrations from Gears

Gears are widely used to transmit power from one shaft to another, usually with a change in speed and torque which makes them vital components in a wide range of industrial and transport applications. They have different configurations where the power transmission shafts can be either parallel or perpendicular to each other. The majority of the gears have conjugate profiles allowing that kinematically, a constant output speed is obtained for a constant input speed. The most used tooth profile is the involute, making the speed ratio almost insensitive to small variations in centre distance [1]. Despite this, during meshing, when a pair of tooth lose contact with each other, a sudden decrease in the torque transmission occurs which set the gears into torsional vibration.

The load applied in the transmission lead the tooth to deform, introducing a meshing error or Transmission Error (TE) even if the tooth profiles were perfect. In addition to this, there are, both intentional and unintentional geometric deviations from the ideal profiles. $Tip - relief$ is a typical tooth correction that allows each tooth to come into mesh without impact, reducing the vibration and noise for a particular load. The TE is a function of tooth deflection, which varies with the load. This makes the amplitude of the resulting vibration at

the toothmeshing frequency, varying directly with the load fluctuation when in service, and may be considered as an amplitude modulation effect. Despite it load-dependent component, transmission error is also a function of geometric deviation from the ideal profile. There are three different types of TE that can be distinguished.

- *Unloaded Static* - is measured when a test gear is meshed with an ideal master gear under a very light load, only enough to keep the teeth in contact.
- *Loaded Static* - includes the tooth deflection due to a constant load torque and can be measured under slow roll conditions.
- *Dynamic* - is the one in operation, where dynamic effects cause fluctuations in the torque transmitted by the gear set.

The characteristic of a gear vibration signal is the presence of a family of sidebands around the gear meshing frequency and its harmonics. In order to identify a family of sidebands, analysis in the cepstrum domain is a well-established technique which has been applied widely in CM programs. The effects of the gearmesh frequency and its harmonics can be present in all tooth pairs or in singular pairs, which will induce a variation that repeat for each rotation of each gear. The last ones manifest at the harmonics of each gear rotational speed and sidebands on the gearmesh frequency. The gearmesh frequency and harmonics are due to tooth deflection, profile deviations or uniform wear over the teeth, which is the one enabling CM of gears to detect faults. The harmonics distribution and sidebands provided the differentiation between faults. It is relevant to remember that there is some signal absorption between the fault source and the measured response due to the transfer path and respective transfer functions. Analysing the spectrum, Randall [1] propose the diagnostic detailed on table 2.1.

Table 2.1: Harmonics and sidebands interpretation

Cause	**Spectrum**
Slow Variation	Runout, distortion, non-uniform wear. Low harmonics and sidebands around toothmesh are affected
Local Faults	Tooth root cracks, spalls. Large harmonics distribution and sidebands results
Random Errors	Random tooth spacing error. Large harmonics distribution and sidebands results
Systematic Errors	Ghost components from the gear cutting machine

For spur gears, the gearmesh frequency is the product of the number of gear teeth and the rotational speed [7] as seen in equation (2.29) where N_P and N_G are the rotational speed of the pinion and gear in rpm, and T_P and T_G are respectively the number of teeth.

$$\text{GMF (Hz)} = \frac{N_G T_G}{60} = \frac{N_P T_P}{60} \tag{2.29}$$

For planetary gears, where multiple output speeds can be obtained, the gearmesh frequencies between sun-planet GMF_{SP} and planet-ring GMF_{PR} are given by equations (2.30)

and (2.31), where N_S, N_R and N_C represent the rotational speed, in Hz, of the sun, the outer ring and the carrier respectively, while T_S and T_R are the number of teeth in the sun and outer ring respectively. The carrier speed can be calculated using equation (2.32).

$$\text{GMF}_{SP} = |N_S - N_C| \cdot T_S \quad \text{(Hz)} \tag{2.30}$$

$$\text{GMF}_{PR} = |N_R - N_C| \cdot T_R \quad \text{(Hz)} \tag{2.31}$$

$$N_C = \frac{N_R \cdot \frac{T_R}{T_S} - N_S}{\frac{T_R}{T_S} - 1} \tag{2.32}$$

2.2.3 Generated Signals

2.2.3.1 Unbalance

The most common type of interference in the signal is unbalance which occur in every rotating machine even if with a residual value [11]. This is due to an uneven distribution of the mass on a rotating component about its axis which creates an unbalance force that acts radially being proportional to the square of the angular rotational speed and the radial displacement between the centre of mass and the centre of rotation [12]. The system response depends on the inertia distribution and on the location of the speed relatively to the resonance frequencies. The additional forces generated by unbalance induce additional fatigue load on the bearings and lead to premature failure. The main causes associated with unbalance are different materials density, imperfections occurred during manufacturing, asymmetrical shapes, assembly errors or even thermal distortion of the shaft during running.

The vibration amplitude can be interpreted as the severity of the unbalance [11]. The system response to the unbalance can be understood using the simplest case called Laval-Jeffcott, which has a single concentrated disc symmetrically located in the middle of the flexible shaft. Despite the distributed mass and elasticity along the shaft which allows an infinity of Degree of Freedom (DOF)'s, if we consider that the mass of the disc mounted on the shaft is much higher than the mass of the shaft, we can treat the system as discrete presupposing the mass of the shaft negligible and the lateral stiffness of the system as the lateral bending stiffness of the shaft [12]. On real life situations, the shafts are usually axysimmetric and the bearing housings have different stiffness in horizontal and vertical directions, thus the response will be different in two directions. In addition to that, the bearing stiffness is usually nonlinear because of the fluid film that supports the load. This nonlinearity will lead to higher harmonics and can generate axial responses even in cases where only radial unbalance is present.

For simply comprehension, lets consider the system as isotropic which means that the system stiffness is the same in the two direction x and y. The response amplitude and phase for the isotropic system are given by the equation (2.33a) and (2.33b) [12], where $r = \omega/\omega_n$ represent the ratio between the angular speed and the natural frequency, ξ the damping ratio, e the distance between the centre of mass and the centre of rotation, m_0 the unbalanced mass and M the machine mass.

$$X = Y = \frac{m_0 e}{M} = \frac{r^2}{\left[(1-r^2)^2 + (2\xi r)^2\right]^{1/2}} \tag{2.33a}$$

$$\phi = \tan\left(\frac{2\xi r}{1-r^2}\right) \tag{2.33b}$$

The variation of the amplitude with the ratio r and the damping ration ξ is shown in the figure 2.4.

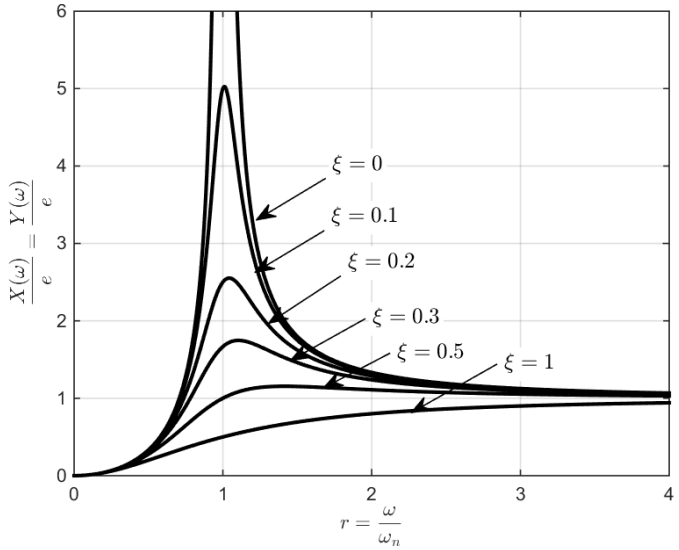


Figure 2.4: Amplitude of the response with the frequencies ratio r and damping ξ

For angular speed below the natural frequency, the movement is controlled by the stiffness. For angular speeds above the critical there is the tendency for the disc to rotate about its centre of mass which means that its trajectory will be inside the one of the geometric centre. This effect is known as self-centering and allows the forces induced to the bearings to be reduced because they are only proportional to e instead of $e\omega^2$ as for speeds below the critical. To surpass the resonance frequency on the start, the system is rapidly accelerated not allowing the amplitude to increase to infinite. There are standards which establish limits for unbalance according to the applications. The ISO 1940 is one of the most widely followed and defines the quality grade G as:

$$G = e \cdot \omega \tag{2.34}$$

There are nine balancing grades that are suitable for different situations [7]. Some of them are detailed in the table 2.2. The most strict one, used in gyroscopes and precision equipment is G0.4 and the most rough is G630, used for crankshafts of rigidly mounted diesel engines. Typically the values used are G2.5 and G6.3 [11]. Other standardization is the American Petroleum Institute (API) with parameters around 4 times more stricter than the suggested by the ISO.

Table 2.2: Some ISO1940 quality grades and applications

Quality Grade	Application
G4000	Crank shafts of rigidly mounted slow marine diesel engines
G250	Cranj shafts drive of rigidly mounted fast four-cylinder engines
G6.3	Processing-plant machines, turbine gears and impellers
G0.4	Spindles, discs and gyroscopes

2.2.3.2 Misalignment

Misalignment occurs when two shafts, driver and driven, are not in a straight line or when there is an angle between them, which primarily happens from differences in the elevation level [7]. The misalignment can be, as shown in figure 2.5 parallel, meaning that one of the two shafts is displaced laterally, keeping parallel to each other or angular misalignment where the axis of one is at an angle to that of the other. As unbalance, misalignment can lead to premature failure due to the bending deflections introduced in the shaft. These induced moments have to be counteracted by forces at the bearings and the foundations.

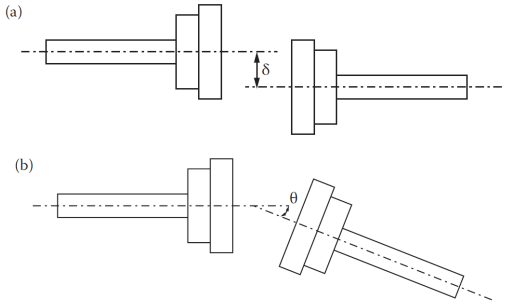


Figure 2.5: Parallel(a) and angular(b) misalignment [7]

The misalignment can be internal, corresponding to the co-axiality of the bearings with respect to each other, or external when two machines are coupled together. The internal alignment is done by bearing housings, casings or supports and is assured during the equipment design. The alignment of the shafts should be done under operating conditions and not stationary to take into account the thermal expansion of the supports and deformations on the structures. This can cause some machines to run misaligned while cold. Due to misalignment, there is an axial push and pull on the shafts and a strong axial vibration is detected. This problem can be easily detected in shafts by the increase of the axial vibration levels at a frequency twice the rotating speed of the shaft. This is explained by the fact that an increase in the misalignment can cause the torque to increase at twice the rotational speed according to [7]. Usually flexible couplings are used to eliminate the effects of misalignment which can also introduce vibrations depending on their own properties. The classical universal coupling is the Cardan joint. This induces a torsional vibration into the driven shaft, with relative instantaneous velocity, given for the input shaft rotation angle θ by the equation (2.35), where α represents the misalignment in degrees.

$$\frac{\omega_2}{\omega_1} = \frac{\cos\alpha}{1 - \sin^2\alpha \cdot \sin^2\theta} \tag{2.35}$$

The previous equation goes through two periods of oscillation for each rotation of the shaft as can be seen in figure 2.6. The bending moments, providing from the torsional vibrations, will generate a torque $T\cos\alpha$ and bending moment $T\sin\alpha$ on the driven side bearings, for a torque T on the input shaft.

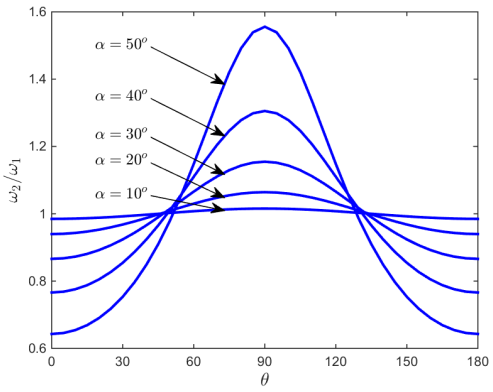


Figure 2.6: Torsional vibration in a Cardan joint

A study conducted by [1] on a gearbox between a gas turbine running at 5100 rpm and a synchronous generator at 3000 rpm before and after realignment result in a reduction of 14 dB in the vibration level of the first harmonic on the generator shaft and 8 dB on the first harmonic of the turbine shaft. Table 2.3 provides the residual misalignment values considered safe based on the machine operational speed.

Table 2.3: Alignment tolerances according to rotating speed [11]

Speed(rpm)	Offset (μm)	Angle (mrad)
<500	127	1.5
500-1250	102	1.0
1250-2000	76	0.5
2000-3500	51	0.3
3500-7000	25	0.25
>7000	13	0.2

2.2.3.3 Calibration Solutions

While working, a shaft that experience unbalance and misalignment will acquire a permanent bow. These deformation can lead to the development of a crack, which is a common fault detected in large rotors such as turbogenerators. Currently, there are several methods to mitigate the problems due to machine looseness.

Single plane balancing is a balancing method used in machines that operate below their natural frequency and have a L/d ratio lower that 0.5, where L and d represent the rotor length and diameter respectively. This method is not advisable for speeds greater than 1000 rpm. For ratios between 0.5-2 the method can be applied for speed up to 150 rpm but other techniques will return better results. In machines with L/d ratios higher than 0.5 a two-plane balancing method should be applied. However, in situations where the machine runs above the

natural frequency, the number of planes that should be used for balancing is $N + 2$, where N is the number of resonance frequencies below the operating speed [11]. A single plane balancing method has very limited application due to the geometry and speed restrictions. This method is applied using a mass with the purpose of create a new unbalance and a stroboscope to measure the mass phase. After two measurements, with and without the new mass it is possible to know the location and weight of the mass in order to reduce the unbalance. When adding more balancing planes, cross-effects may occur. Those are characterized by misinterpreted unbalance at one end of the shaft that can actually be created at the opposite end of it. This will make unbalances appear in planes which are not their respective balancing ones.

Several conventions use a dial indicator during the alignment process. The dial gage is attached to a bracket and fixed in one end of the shaft. The methods are the two dial alignment which uses four measures and is done in the vertical and horizontal plane. The dial can be set to measure parallel or angular misalignment. Using a three-dial method of alignment the required time for readings is considerably reduced and the two types of misalignment are measured simultaneously, avoiding dial reposition and reconfiguration. Despite this fact, this method is not accountable for axial floats, as floating shaft engines can experience up to 10 mm of axial displacement. It is also possible to proper align machines using the reverse dial method which is generally employed on dual articulation coupling. This method is not affected by axial movements of the shafts while the geometric accuracy is better than the two-dial method. In addition to this, the couplings do not have to be opened for measuring and thermal corrections can be incorporated. There are other sensors that can be used for shaft alignment such as lasers which minimize the error and the time consumption. According to [11] there are laser systems which need less than a quarter turn of the shaft to produce very accurate results. In both cases, the appropriate method is based on the type of machine, rotational speed, the maintenance policy and tolerances. The machine importance in production will determine the amount of time that should be assigned for calibration.

2.2.4 Signal Classification

In machinery condition monitoring, decisions are made based on the nature of the measured signals. Those signals carry information and their features can be constant or variant on time. The majority of the machine components are characterized by specific vibration signals, that can be repetitive on time, which allow them to be separated from others, such as the random interferences, in order to perform different analysis and distinguish fault from healthy conditions. For instance, the characteristic frequencies of rolling element bearings are generally not harmonics of the associated shaft speeds while gearmesh signals are at harmonics of the associated shaft rotation speeds [1]. Signal processing is simply the manipulation of the properties of a specific signal to obtain another one with preferable properties. In terms of continuity of the independent and dependent variables, signals can be classified in four different ways which are now explained [13].

- *Analogue* - Both independent and dependent variables defining the signal are continuous in time and amplitude, in other words, for each specified time instance the signal as a specified amplitude value.
- *Continuous-Time* - The time variable is continuous in the range where the signal is defined. If the variable in representation is x and the time variable t the signal is expressed as $x(t)$.
- *Discrete-Time* - Time variable is discrete in the range for the signal definition If the discretization of time is done using n values then the signal is expressed as $x(n)$. However, the amplitude of a discrete-time signal can take any value inside a specific amplitude range which make the first a continuous variable.

- *Digital* - Occurs when the amplitude of a discrete-time signal is also discrete.

Vibration signals can be divided into various categories and classified according to those. The type of signal generated by a component determines the processing which should be applied to it. Signals are usually distinguished by the repetition of periodic events therefore, one of the most basic ways to evaluate signals, is in terms of their frequency spectrum which allow one to percept how their constitutive components are distributed with frequency [1]. To perform the transformation between time and frequency domain Fourier analysis is commonly used and will further be detailed in this text. The primarily division in signal classification is between stationary and non-stationary as can be seen on the figure 2.7. Signals whose statistical properties do not change with time which means that are invariant in time are defined as stationary signals. Deterministic signals are essentially composed from discrete frequency sinusoids and therefore their frequency spectrum consists of prominent discrete lines at the frequencies of those sinusoids. Signals from rotating machines operating at a constant speed are examples of stationary deterministic signals [7]. When a signal is deterministic and after their initial conditions are known, the value of the signal can be *determined* at any moment of time, future or past.

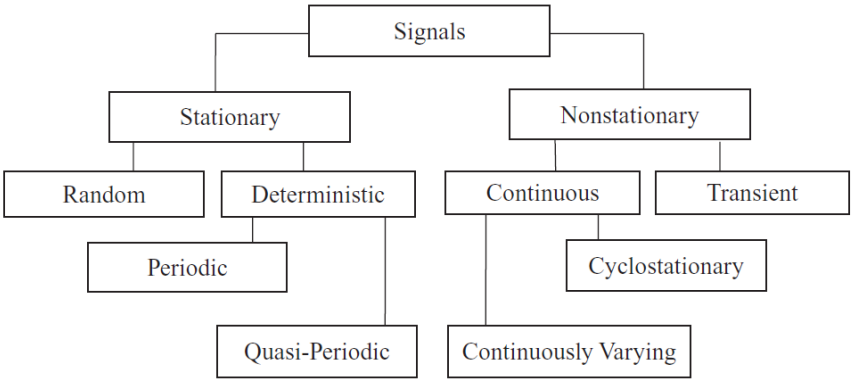


Figure 2.7: Signal classification [1]

Random signals cannot be predicted at any time which make them very complex in terms of analysis. Many real life signals are random for instance the noise produced by rain drops on a roof. However, for stationary random signals, their statistical properties are unchanging with time. The statistical properties can be obtained by averaging a considerable number of measurements. If the function being averaged using the expectation operator $E[\cdot]$ is the signal itself, then the result will be the mean value. If the signal is squared, then the result will be the mean square value. This is valid if the signals are stationary and ergodic. The term ergodic mean that the statistical properties can be deduced from a single, sufficiently long random sample of the measures. As an example lets took the vibration signals measured on a number of vehicles driven at constant speed around a uniform track. If we use small and large cars, it is possible for the process to be stationary but for it to be ergodic, all the vehicles would need to have the same size [1].

Non-stationary signals statistical features change in time domain and can be further divided into continuously or transient. In general, transient signals can be described as those with finite length on time. The most common type of transient signals are due to an impact for instance the sound of a slamming door. These signals can also be repetitive, like a jack hammer. Cyclostationary signals have the advantage of varying periodically with time therefore their total energy is finite. Non-stationary vibration signals, measured during the run-up or

coast-down of a machine also have a finite length, but are likely to be considered as continuously changing. Those can be divided into short *quasi-stationary* sections, to perform analysis in the time-frequency domain.

2.2.4.1 Stationary Deterministic Signals

As was already mentioned, deterministic signals are composed of discrete frequency sinusoids. The two types of deterministic signals periodic and quasi-periodic vary on the *"harmonics"* of the Fundamental Periodic Frequency (FPF). For periodic stationary deterministic signals, the components appear in the spectrum as integer multiples of the FPF while on the quasi-periodic the discrete frequencies do not totalize an harmonic series. In other words, this means that the frequency ratio is not an integer, as happens for perfect harmonics, but an irrational number [1].

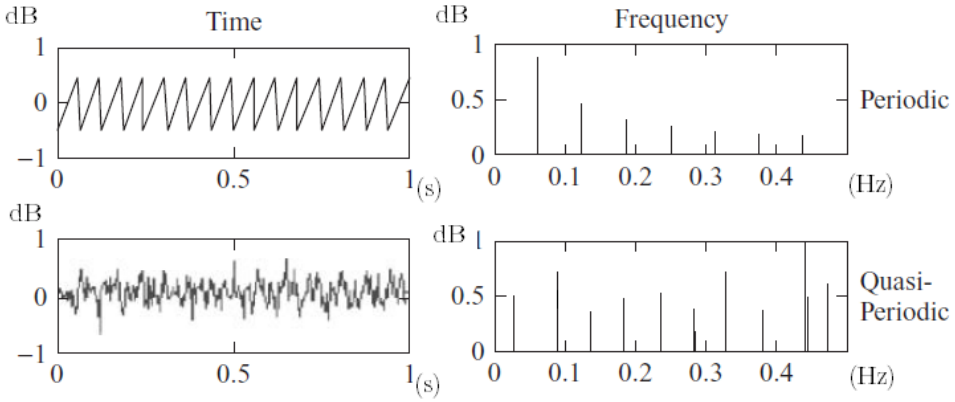


Figure 2.8: Time and Frequency domain of deterministic signals [1]

A typical example is the vibration from a gas turbine with several independent shafts. Despite every single shaft generate a family of harmonics while running, the combination of all the shafts will be a quasi-periodic signal made of periodic sine waves. According to Randall, a quasi-periodic signal can be treated as a special case of a periodic signal where the period tends to infinite. Figure 2.8 represents the time and frequency domain for both periodic and quasi-periodic signals.

2.2.4.2 Stationary Random Signals

Stationary random signals have a similar time signature as the quasi-periodic ones but their spectrum allow to understand the differences since the first one does not have discrete frequencies and has a range over the frequency domain. One example of random signal is displayed in figure 2.9. The properties of a random signal must be described by their statistical parameters such as mean value or mean square value. The equations for the mean value and mean square value are those following, where E represents the expectation operator.

$$x_m(t) = E[x(t)] \tag{2.36}$$

$$x_{ms}(t) = E[x^2(t)] \tag{2.37}$$

The mean square value can be worked to have the same units as the mean value using the root of the mean square value. This feature is called Root Mean Square (RMS) and will be discussed in the following chapters among other features. One other feature that can be applied for random signals is the autocorrelation function which overlaps the signal with itself but displaced on time in order to understand how well a signal correlates with itself. The common definition of autocorrelation is given by equation (2.38).

$$R_{xx}(t,\tau) = E[x(t-\frac{\tau}{2}) \cdot x \cdot (t+\frac{\tau}{2})] \tag{2.38}$$

Since stationary random signals are independent on time t the autocorrelation function can be written as a function of time displacement τ.

$$R_{xx}(t,\tau) = R_{xx}(\tau) \tag{2.39}$$

If it is assumed that the signal is ergodic, as explained before, the averaging of the measures can be performed on a specific long enough length of a single record and the equation (2.38) can be rewritten as (2.40).

$$R_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x \cdot (t-\frac{\tau}{2}) \cdot (t+\frac{\tau}{2}) dt \tag{2.40}$$

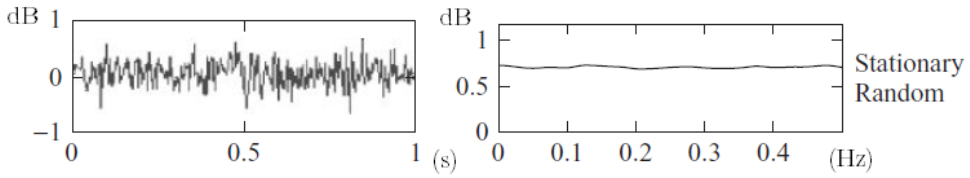


Figure 2.9: Time and Frequency domain of discrete random signals [1]

2.2.4.3 Cyclostationary Signals

In the real world, several processes derive from periodic phenomena and although not periodic on time, generate random data whose statistical characteristics vary periodically with time and are therefore designated cyclostationary processes [14]. The signals generated by rotating machines are cyclostationary. Some examples of mechanical and acoustic signals with some approximate periodicity are for instance the rotating blades of a wind turbine, rolling element bearings or engine pulleys. Cyclostationary signals are characterized by an amplitude modulation of the cyclic signal which results in pairs of sidebands in the spectrum, spaced around each modulated frequency component by an amount equal to the modulation frequency [1]. The biggest problem in rotating mechanism diagnosis is to identify the modulating frequencies that contribute to the cyclostationary nature of the signals.

In telecommunications, telemetry, radar and sonar applications, periodicity is due to modulation and coding operations. In mechanisms as had been said is due to rotation. In astronomy, the periodicity is originated by revolution and rotation of planets, in econometrics results from the seasonality. The first studies in cyclostationary signals were conducted by William Gardner in 1985. Until then, the classical signal analysis was based on the assumption of stationarity [15] which represented a problem for rotating equipments which hardly fit

conditions of stationarity. The most useful parameters in signal analysis are the frequencies characteristics, therefore the stationary approach, which assume time-invariant statistical parameters could not be applied. Due to this fact, Gardner started to develop a appropriate probabilistic and statistical framework that he applied in the context of telecommunication signals.

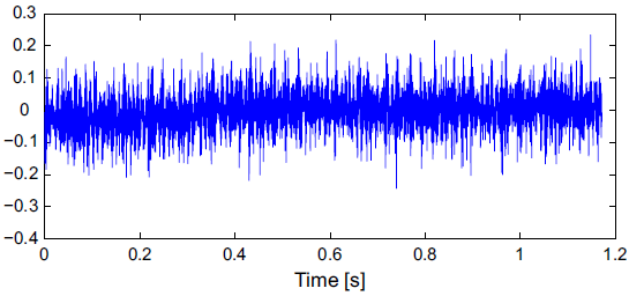


Figure 2.10: Example of a cyclostationary signal generated by rolling bearings [15]

The signal represented in the figure 2.10 is from a rolling elements bearing and a randomness can be observed. In fact, assessing the cyclic nature of a signal is an added difficulty. Although on some cases, periodic patterns can be observed, these become more clear after synchronous characteristics are calculated. The analysis on the frequency-domain for nonstationary signals is very technical and complex and some authors overpass the problem by choosing an appropriate window where the signal is considered stationary. However, this assumption is not admissible due to the fact that all the subsequent analyses will depend on the window choice [15]. Cyclostationary models as we will see ahead use a two-dimensional spectrum instead of the usual spectrum enhancing the frequency signature of the signal. A stochastic model is applied to identify cyclostationary signals. These models, use the autocorrelation function, seen in equation (2.39) that vary periodically with time and that can be expanded in a Fourier series whose coefficients depend on the lag parameter τ. There are signals, designated quasi-cyclostationary which occur when the frequencies of the Fourier expansion are not proportional to the period of cyclostationarity as happens for perfect cyclostationary signals [14].

Assuming that a signal $\{X(t) : t \in Z\}$ and that Z represents the integers, the cyclostationarity of $X(t)$ is a repeatable behaviour for the first and second order characteristic of the signal. In other words, the signal will be called cyclostationary if the mean, given by equation (2.36) and the autocorrelation function, given by equation (2.38) are periodic or *quasi-periodic* in t. When the exact period P of the autocorrelation function $R_{xx}(t, \tau)$ is known, the multiples of the fundamental frequency α can be calculated by equation (2.42) called cyclic correlation.

$$\alpha = \frac{2\pi}{P} \tag{2.41}$$

$$C(\alpha, \tau) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} R_{xx}(t, \tau) \tag{2.42}$$

The estimator E used to calculate the mean and variance, should minimize the probability of false alarm and detection. Usually, this was accomplished by assuming certain signal distributions such as Gaussian or large-sample properties of the estimator. In the context of vibration signals, the assumption of gaussianity cannot be verified due to lack of stationarity and

large deviation of the signal, therefore the estimator of the cyclic correlation is approximated by a Gaussian distribution.

Chapter 3

Bearing Diagnosis

Bearing Analysis

Nowadays, bearings have a critical role in machinery, both domestic and industrial. Therefore, the proper functioning of these equipments depends on the smooth and quiet running of the bearings. An early fault detection may prevent catastrophic failures or malfunction. Rolling element bearings are widely used in rotating equipment due to their relatively low price and operational facility. However, these components are responsible for almost 45-55% of the machine failures [3], due to the presence of defects such as galling, spalling, peeling, subcase fatigue or due to installation problems such as misalignment and unbalance which was already been addressed in section 2.2.3. Defects in bearings may arise during the running or the manufacturing process which make them, important objects of analysis in CM. This section, will focus on the diagnosis and prognosis of bearing defects using vibration methods. Other methods could also be applied as discussed in the chapter 2.

The generation of vibration and noise in bearings is a well studied topic [16]. Those studies concluded that bearings act as a source of vibration due to either varying compliance or the presence of defects, even if we could assume that they are geometrically perfect, which is not the case as discussed before. This is due to the number of rolling elements used to carry the load is finite and their position in the load zone change with the shaft rotation which cause stiffness variations and consequently vibrations. According to [16], when the bearing races are assumed as continuous systems, the changing direction of the contact forces applied by the rolling elements may cause flexural or ring-mode vibration of the races. Despite all those vibrations in the bearings, the presence of defects increase significantly the level of them and allow fault detection.

Returning to figure 2.2, which represents the vibration signals at each component, it will now be detailed how those signals are generated. A fault on the outer race of the bearing, localized on the load zone, generates a pulse on the waveform every time a rolling element passes over the defect and hits it. This means that if the fault is not in the load zone it could be more difficult to detect because the pulse magnitude will not be notorious. Typically the pulses should have the same magnitude due to the fact that the load conditions must be constant and despite the variations on the element along the shaft rotations, all the rolling elements must experience the same load changes each revolution.

For defects on the inner ring, there is a pulse on the waveform, each time the inner ring passes over a rolling element, considering that the inner ring is the one rotating. It is observable that there are different magnitudes in the pulses for the generated signal. This is due to the fact that when the inner race is rotating and the defects hit the rolling elements, those may not be in the loaded zone which will cause different magnitudes in the signal. In other words, the magnitude will be as high as the load applied to the contact between defect and rolling

element. The inner ring fault signals are often surrounded by sidebands at intervals of about one time the rotational frequency [8]. In rolling elements fault, there is a pulse generation at each contact with inner and outer ring. Once again, the magnitude of those pulses are related to the position of the rolling element, i.e. in a loaded or unloaded zone. The signal is amplitude modulated. The signals acquired by the diverse condition monitoring techniques, provide the data for analysis in a raw way and have to be treated precisely to execute diagnosis and prognosis of rolling element bearings. To provide a better degree of accuracy, more data should be gathered and analysed which can take longer to process.

3.1 Low-Speed Bearings

Currently is possible to find several equipments running under low rotational speed such as paper machines, cooling tower fans, wind turbines, rotary dryers, low-speed agitators, ball mills, crushers or mining equipment. According to [17] and despite no universally accepted criteria for classifying the rotational speeds, it is considered that a machine is operating at low-speed if working under 600 rpm. Other regimes of speed would be medium and high speed. To properly adopt a CBMS to low-speed machines it should be kept in mind that these equipments change in slight increments and have low energy emitted to certain damages which can cause the failures to stay latent until a critical point is reached. Some high speed machines such as gearboxes and compressors may also exhibit faults for the low-speed region and the following considerations stay valid for these equipment.

In low speed equipments, the usage of inadequate monitoring techniques may not be sensitive enough to detect changes at low frequencies. This is a common mistake in some maintenance plans which use the instruments for high-speed machinery monitoring in low-speed machinery. The changes in the equipment condition happen slow and the people could get used to it and not even notice subjective changes, detected by listening or seeing. It is critical to compare the data with the original state of the machine due to these small increments which individually may not represent any notorious change. Typically the fault for these type of equipment develops very slow and the time between the development of a potential failure in an effective one is very long which has the advantage of avoiding some catastrophic and suddenly failure but, on the other hand, may not be very evident at the early stages. The steps for a maintenance strategy are the same for all the other applications and can be resumed in three steps.

1. Data Collection
2. Processing and Analysis
3. Diagnosis and Prognosis

The analysis of low-speed machines differs from other regimes in two main factors. The overall energy generated by impacts are lower than for other regimes and usually, these equipment is installed in heavy duty machines which can create conditions of nonstationarity, making the fault detection more complicated [17]. The considerations for these regime have to be taken into account since the beginning of the maintenance strategy because due to the low energy, the faults can be falsely attributed to noise from the environment. The last paragraph can be easily understood by the figure 3.1

The forces generated at low speeds are very low. To have and idea, if we think about an unbalanced disk rotating with a shaft, the forces at the supports are given by $F = me\omega^2$.

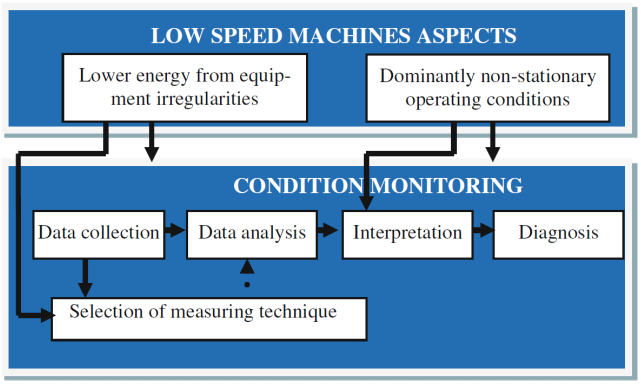


Figure 3.1: Low-speed machines difficulties [17]

This means that a decrease of 10 times the speed will decrease 100 times the forces in the bearings, causing the amplitude of the pulses generated by the faults to be very low, as had been discussed previously for the signal generation explanation.

3.2 Bearing Defects

In the previous section we conclude that bearing defects cause a significant increase in the vibration level. These defects can be divided into distributed and local ones. The distributed defects include surface roughness, waviness, misaligned races and off-size elements [16]. The evaluate the surface conditions, its characteristic parameters are compared in terms of the wavelength with the Hertzian contact width of the rolling element with the races. The surfaces are considered rough when their wavelength is of the order of the contact width. On the other hand, longer wavelength features are considered as waviness. These defects are caused by manufacturing error, improper installation or abrasive wear. The variation in contact force between elements and raceways due to distributed defects, results in an increased vibration level. Localized defects include cracks, pits and spalls on the rolling surfaces [16]. The most common mode of failure for rolling element bearings is spalling of the races or the rolling element bearings which are caused by a fatigue crack below the surface as we will see forward. Fatigue failure is related to overloading or shock loading on the bearings during running or installation. For localized defects, when there is an interaction with a rolling element or a race, depending where the fault is located, there are abrupt changes in the contact which generates a pulse of short duration and the spikes in the time signal.

3.2.1 Distributed Defects

3.2.1.1 Surface Roughness

As had already been said there is a difference between ideal and real surfaces. Real bodies have form errors, surface waviness and roughness due to the manufacturing or simply by previous operation. This means that in contacts between rolling elements and raceways, the real contact area is smaller than the apparent one and the border of the real contact area may expand outside the theoretical one. According to the Hook's law and Poison's ration which declare that when a rolling surface is compressed in one direction, its material tends to expand

in the transverse direction [18]. This creates tangential forces and local slip between contact surfaces at the edge of the rolling surface.

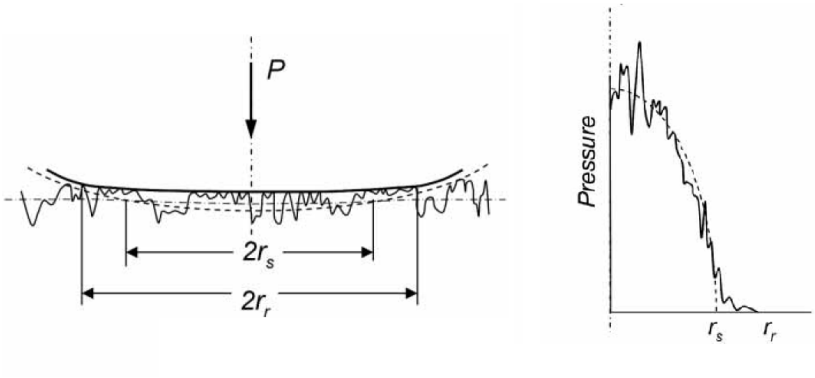


Figure 3.2: Ideal and real Hertzian contact and normal pressure distribution [18]

In the figure 3.2 it is observable the ideally and real contact surface and contact width. The broken line represents the ideally smooth elastic contact, where r_S represents the semi-width of the contact. On the other hand, the solid line represents a randomly rough surface where r_R represents the semi-width of the rough contact. One can observe that r_R is greater than r_S, as explained before. On the right of the figure, it is observable the normal pressure distribution with r.

There are several parameters that allow one to assess the surface rugosity. The amplitude parameters characterize the surface based on the vertical deviations of the roughness profile from the mean line. The most widely used is the arithmetic average R_a of the absolute values. There is also a Deutsches Institut für Normung (DIN) standard that defines the amplitude parameter as the average distance between the highest peak and lowest valley in each sampling length. The equations for R_a and R_{DIN} are given by (3.1) and (3.2) respectively, where n represents the number of samplings and R_t represents the maximum height at each measurement, which is given by the maximum valley depth plus the maximum peak height.

$$R_a = \frac{1}{n}\sum_{i=1}^{n}|y_i| \tag{3.1}$$

$$R_{DIN} = \frac{1}{n}\sum_{i=1}^{n}R_t(i) \tag{3.2}$$

$$R_t = R_{Max+}R_{Min} \tag{3.3}$$

3.2.1.2 Film Thickness

In addition to the surface roughness, another parameter that will largely influence the surfaces wear is the film thickness and the regime of lubrication, which is determined based on the surface roughness. In the Chapter 2 was referred that there is a difference between hydrodynamic and rolling element bearings that differ in a rough way on the lubrication regime while running. The entrapment of a lubricant into a converging gap at a sliding or rolling contact creates a lubricant film which aim to separate the contact surfaces and reduce the contact pressure. The ratio between the lubricant film thickness (h) and the combined surface roughness (R_A) in the contact determines the lubrication mechanism [18] and is from now on designated

as λ.

$$\lambda = \frac{h}{R_A} \tag{3.4}$$

$$R_A = \frac{1}{2}\left(R_{a_1} + R_{a_2}\right) \tag{3.5}$$

The lubrication regimes may be divided into boundary lubrication, mixed lubrication, elasto-hydrodynamic lubrication and hydrodynamic lubrication. In operating conditions where the hydrodynamic action is too weak for separating the contact surfaces, such as in low-speed applications, the surfaces are only wetted by oil. The regime is characterized by $\lambda < 1$ and the load is carried mainly by solid-to-solid contacts. Applications that use boundary condition lubrication regimes rely on lubricant and surface mechanical properties to develop boundary films while running. For rolling element bearings, boundary condition may occur locally in micro-slip zones and at the end planes of axially loaded rollers.

A mixed lubrication regime is characterized by boundary lubricated conditions as part of a elastohydrodynamic or hydrodynamic lubrication. The ratio between film thickness and combined roughness is between $1 < \lambda < 3$. In other words, the contact can be lubricated by a solid or viscous boundary film. The elastohydrodynamic $3 < \lambda < 10$ and hydrodynamic $\lambda > 10$ lubrication regimes are based on Reynold's equation for hydrodynamic oil pressure build-up [18] and a combination of elastic deformation at the contact surfaces of low geometric conformity. The slide-to-roll ratio in a rolling contact influences the oil film generation. In the elastohydrodynamic case the film thickness is slightly higher than the combined surface roughness, while in hydrodynamic regime is much higher. In rolling element bearings, hydrodynamic lubrication occurs in contacts between rolling elements and their cages and also between roller ends and bearing race flanges for roller bearings with axially loaded rollers. The higher film thickness achieved in this regime provides a significant degree of damping for dynamic loads.

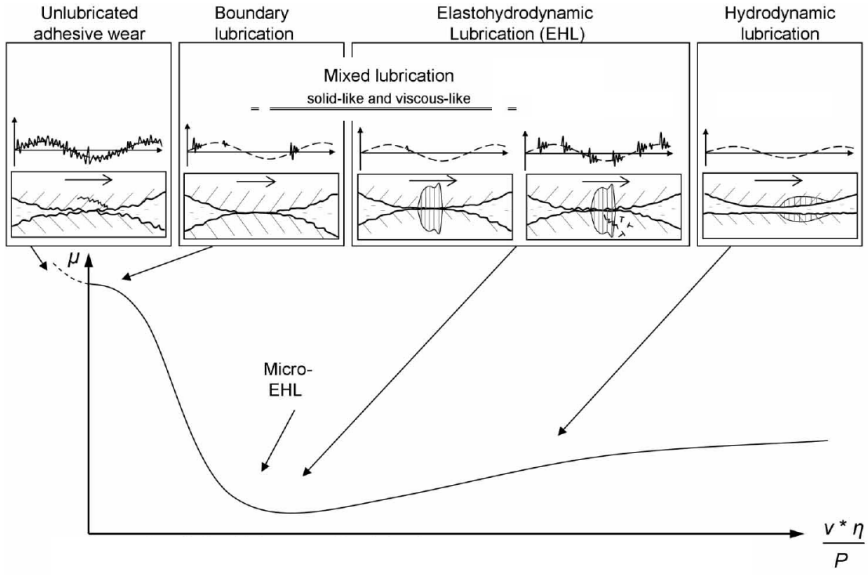


Figure 3.3: Representation of a Stribeck curve and lubrication regimes [18]

From figure 3.3 it is observable the Stribeck curve, where μ represents the coefficient of friction, v the velocity, η the viscosity and P the load [18]. It is also shown the vibration response in the different regimes, where the dashed and solid line represent the torsional and

transient vibration respectively. For the boundary regime, the vibration increase due to adhesive wear, which will be following addressed, and surface roughening, which characterized by random impacts and rotational vibration. For the elastohydrodynamic regime the periodical impacts are reduced exciting essentially high frequencies. For the hydrodynamic regime, the vibration response is dominated by rotational vibration. With the increase of the Reynold's parameter, there is an increase in the damping by the oil film which lead to vibrations reduction.

3.2.1.3 Fatigue Wear

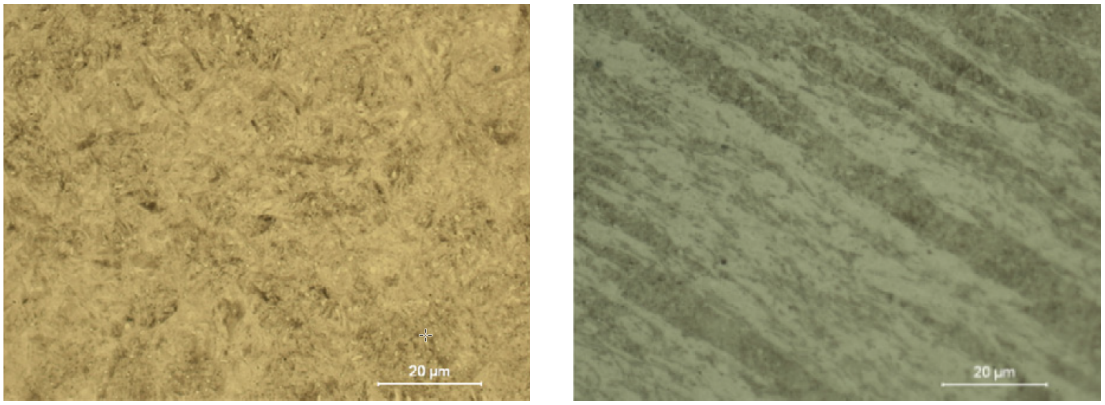
Fatigue wear can occur in surfaces which are dynamically loaded. Fatigue cracks can be generated by diverse mechanisms such as contaminant particles or simply by the dynamic conditions in the bearing, the shear stress and plastic deformations. The initial fatigue cracks are usually inclined to the surface and their consequently intersection leads to pit formation and wear. According to Gabelli [19], the commonly adopted definitions for structural fatigue of machine components can be categorized in three regimes.

- *Low cycle fatigue* - is referred to a regime in which the number of cycles to failure (N_f) is less than 10^5 and the failure is initiated from the surface slip bands.
- *High cycle fatigue* - corresponds to a regime where N_f is between 10^5 and 10^7 cycles. The fatigue failure may be related to surface slip bands or inclusions inside the material.
- *Very high cycle fatigue* - occurs when N_f exceeds 10^7 cycles. It is common for high-strength steels and the fatigue initiation is due to material in-homogeneities such as inclusions.

The life for rolling bearings can be estimated in terms of fatigue according to the design, load speed and lubrication conditions. The bearing life theory has been developed by Stribeck in 1896 and a fatigue limit concept was introduced by Palmgren in 1924. In 2007 it was introduced a new standard, the ISO 281:2007, which includes the concept of fatigue limit stress in the assessment of the fatigue life for rolling bearings, based in the material fatigue data, the rolling bearings endurance test data and the use of factors to predict for large number of stress cycles. The classical fatigue limit stress is defined as the stress below which failure does not occur, even for an infinite number of cycles. This normalization is developed on the basis of a probabilistic survival law for crack initiation in rolling contacts and can be expressed by equation (3.6), introduced by Joannides and Harris in 1985. In the equation, S represents the probability of survival, N the number of stress cycles to failure, e is the Weibull exponent, τ_0 is the maximum orthogonal shear stress amplitude, τ_u is the fatigue limit shear stress, z_o is the depth of the maximum orthogonal shear stress, b is the contact main semi-width, l is the length of the raceway contact and A, h and c are constants.

$$\ln \frac{1}{S} = A \cdot N^e \cdot \frac{(\tau_0 - \tau_u)^c}{{z_0}^h} \cdot b \cdot z_0 \cdot l \tag{3.6}$$

The fatigue limit stress is defined in terms of the maximum Hertzian pressure (σ_H) of the rolling contact and nowadays, with high quality steels and manufacturing practices, its value is standardized to approximately 1500 MPa. Figure 3.4a represents the typical aspect of an undamaged microstructure for the outer race of a bearing subjected to 2.3×10^{10} stress cycles with an Hertzian pressure of 2.3 Gpa. On figure 3.4b, an example of extensive microstructure decay on a bearing that run up to 2.3×10^9 stress cycles with an Hertzian pressure of 3.3 GPa, obtained from the Barden bearing manufacturer.



(a) Undamaged microstructure (b) Damaged microstructure

Figure 3.4: Bearing outer race microstructure [20]

3.2.1.4 Adhesive and Abrasive Wear

Rolling bearings may suffer from adhesive wear at roller ends and in micro-slip zones due to poor lubrication conditions. In these sliding contacts, exists a strong adhesive junction between the surfaces in contact due to frictional overheating. This adhesive wear leads to the generation of particles which will contaminate the lubricant. These particles can cause local stress peaks and shorten the life of the bearing due to their mechanical properties and hardness. According to [18], even particles smaller than the film thickness may cause abrasive wear which may take place in contacts between rough surfaces acting as micro-abrasive particles. For example between the rolling elements and their cages, the adhesive and abrasive mechanisms, lead to pitting and spalling, which are localized defects and reduce the expected lifetime.

3.2.2 Localized Defects

3.2.2.1 Cracks

The most common cause for the formation of cracks in the bearings is rough treatment when the bearings are being mounted or dismounted. However other causes such as excessive interference, excessive load, shock load or heat generation may lead to cracks. As a correction to interference, the user can check the load condition and improve the mounting method, according to the bearing manufacturing NSK.

3.2.2.2 Pitting

The pitting phenomena is characterized by deep craters in the surface of the bearing component and are a result of the fatigue cracks generated in the subsurface. Those are a result of the fatigue process and propagate from the subsurface to the surface originating the release of material from the structure. These particles will forward be the cause for abrasive wear. Beside the fatigue process, those may also be connected with moisture in the lubricant for example poor lubrication.

3.2.2.3 Brinelling

Brinelling occurs when loads exceed the elastic limit of the ring material. The marks resulting from brinell are indentations in the raceways which increase bearing vibration. The cause for these type of failure are static overload or severe impact. To minimize the brinelling, careful handling and installation practices should be adopted. Figure 3.5 show on the left a crack fault, in the middle several pits and on the right the brinelling effect on the bearings, obtained from the Barden bearing manufacturer.

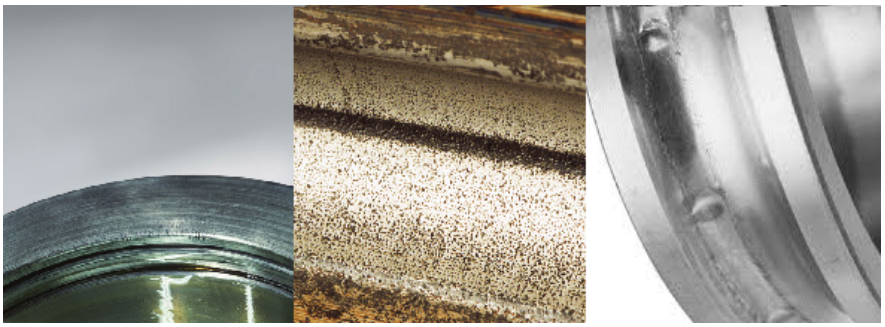


Figure 3.5: Example of bearing defects [20]

3.3 Data Acquisition

Acquiring the correct data is the most important step to evaluate the bearing condition. The condition monitoring has many techniques that were already described and despite their differences, all of them start with data collection. The machine internal geometry has a great influence in the place where the transducer is placed. According to Taylor [21] data should be taken where the transfer function is best, for instance on a bolt head or cover. When dealing with low-speed machinery other considerations arise and it is of primary importance to understand the physical nature of the machinery components function, potential failure causes and the best monitoring parameters. For measuring dynamic effects, like in vibration monitoring, special attention should be taken on the measuring instruments and data collected [17]. The value used to specify the low-speed regime at 600 rpm is due to the representation at which the alert levels relative to vibration must decrease to maintain the security and smoothly running of the equipment.

Low-speed equipments normally require measurements within 0.1-10 Hz bandwidth or even lower than 0.1 Hz. As a matter of fact, motion below the 600 rpm produces very little vibration in terms of velocity but very large displacements. The most important topics to be considered in this regime of speed are the measuring probes, the sampling rate and the filtering of the data.

3.3.1 Measuring Sensors

There are many different types of sensors and manufacturers, providing different options to measure vibration and each one of them has their own advantages and disadvantages. In

general two main type of probes that can be used on condition monitoring, those are accelerometers and displacement probes. At the end of this section, there is a simple explanation on the construction solution and operational principle of the accelerometers, because those sensors are the ones used for the results shown in Chapter 6. Despite this fact, the analyst could adopt other type of sensor to perform the measurements. Figure 3.6 represents the criteria for monitoring dynamical effects according to the range of speeds [17].

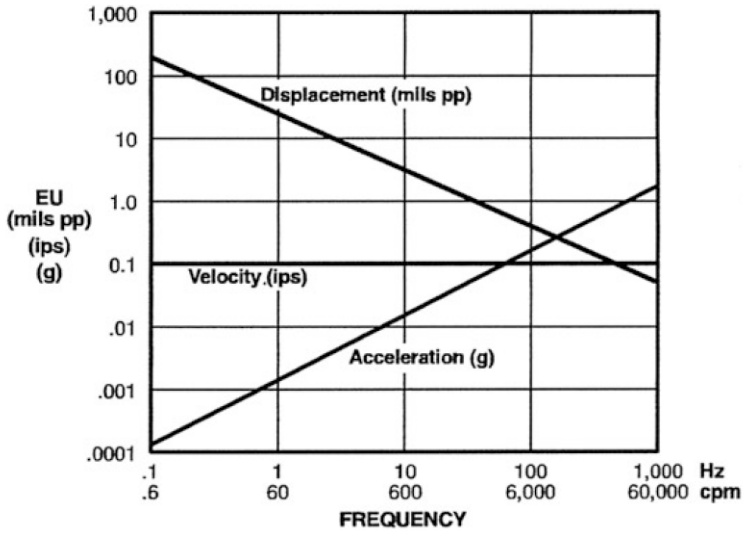


Figure 3.6: Selection criteria for monitoring dynamical effects [17]

The accelerometer, measures acceleration and despite the techniques used for acceleration sensing such as piezoelectric, piezoresistive or inductive, the most important difference is based on the signal response. The accelerometer can be Alternated Current (AC) or Direct Current (DC) type. The first one cannot measure static acceleration but is appropriate for dynamic tests. It is also useful for calculation of velocity and position by integrating the acceleration signal. The most common AC accelerometers are equipped with piezoelectric elements. On the other hand, DC type accelerometers can sense constant acceleration such as gravity which make them able to measure vibrations down to 0 Hz theoretically. The most used DC type accelerometers are capacitive Micro Electro-Mechanical Systems (MEMS) and piezoresistives [17]. The most important characteristics of the accelerometers which can lead in choosing the best option for measures are forward described.

- *Sensitivity* - is the ratio of change in the input, or acceleration, to the change in the output signal. Sensitivity is specified at a particular supply voltage and is typically expressed in units of mV/g.

- *Frequency Range* - the range is usually given in units of gravity constant. For reliable CM the transducer shall be capable of covering a wide frequency range in order to measure bearing resonances and harmonics. The frequency measured should not be greater than the maximum transducer linear range according to the ISO 13373.

- *Resolution* - is the smallest amount of acceleration which can be indicated by sensor.

- *Output Signal* - there are different outputs such as voltage, current output, charge output and Integrated Electronic Piezoelectric (IEPE) which is a piezoelectric accelerometer with a built-in amplifier. The output can also be analogue and digital.

- *Number of Axis* - there are single, double and triaxial accelerometers. The three axis accelerometer will measure acceleration in three orthogonal directions.

In cases such as low-speed conditions, low frequency measurements are required and MEMS sensors are widely used. Those can measure close to 0 Hz and with high accuracy. The typical sensitivity of these sensors begin with 1 mV/g.

The proper attachment of the transducer is critical to accurately transmit the motion of the machinery to the sensor. The methods used for accelerometers are fully described on the ISO 5348. The preferred method, according to ISO 13373 is a rigid mechanical fastening which is commonly achieved by drilled and tapped holes in the transducer and the machine. This connection has the ability to transfer high frequency signals with little or no signal loss. If is impractical to perform a stud-mounted mechanical connection, cements are used to fix the transducer. Those should have high stiffness characteristics when cured. Another technique is to used a permanent magnet, but in order to effectively use magnets, the flatness of the mounting surfaces is imperative. Both cement and magnetic methods could be subject to limits of frequency, temperature and amplitude and therefore should be used with caution according to the same standard referred before. Figure 3.7 represents the location and axis orientation recommended by ISO 13373. This is an option and due to machine geometry or presence of other equipments may not be admissible.

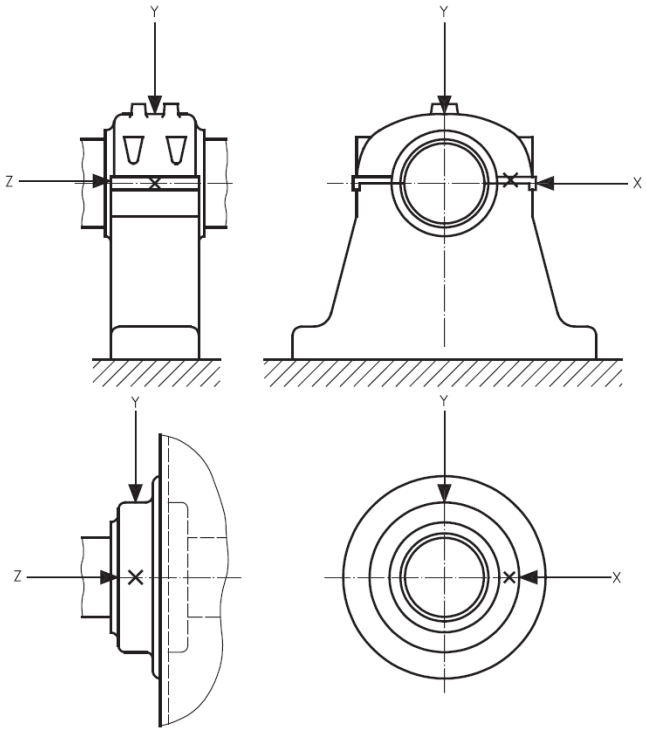


Figure 3.7: Schematic diagram of typical transducer locations [22]

Despite not being used for the data analysed in this text, displacement sensors have, similarly to the accelerometer, different types and solutions. The most important distinction are between contact sensors which operation is based on the displacement change sensing by the direct contact of the sensor with the measured object and noncontact sensor where there is no direct contact between the sensor and the measured object. The first ones are sensitive to wear and the quality of the measured surface while the seconds are robust and use electrical or magnetic field change phenomenons to sense displacement. The noncontact sensors can be capacitive, magnetic-inductive, laser or confocal sensors.

Accelerometers

An accelerometer is a transducer which produces a signal proportional to acceleration. Vibration can be expressed in three parameters which are displacement, velocity and acceleration. A transducer converts one form of energy into another. Ideally these devices should have high dynamic range, high repeatability, low noise and a low hysteresis. The most employed transducers in CM are the piezoelectric accelerometers. This sensors use the properties of certain crystals to generate an electric charge proportional to strain which can after be recorded by a data acquisition system. In the traditional solution, known as compression type, the piezoelectric elements are sandwiched between a mass and the base, which is compressed by a spring. The working range of an accelerometer is localized below the natural frequency of the suspended element. The overall aspect as well as the working range can be seen in the figure 3.8, obtained from the PCB Piezotronics company.

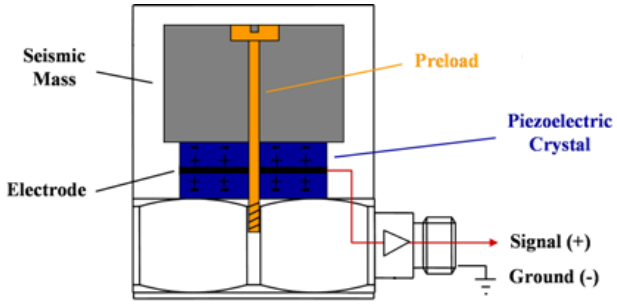


Figure 3.8: Typical design of a "*compression*" type accelerometer

Because the accelerometer is connected to the equipment being measured, the mass is forced to follow the motion of the base by the crystals which act as a very stiff spring [1]. The varying inertial force of the mass, causes a slightly deformation on the piezoelectric elements that translate in a strain proportional to the variation in acceleration and consequently an electric charge proportional to the acceleration. There are also other designs such as the *delta shear*, designed and patented by Brüel & Kjær in which the piezoelectric elements deform in shear. In this configuration the mechanical properties of the assembly are isotropic with no preferential direction. Some other types of solutions allows varying resonance frequencies in different directions.

The electric circuit for data acquisition has very high impedance and is commonly affected by a number of problems such as electromagnetic interference. To minimize this effects one can use coaxial cables with an outer braided wire shield [1]. Also the connectors used for the sensor have a great influence in this topic, for laboratory measurements the connection called *microdot* gives the best accuracy however for industrial application where robustness is a requirement, $double-shielded\ microdot$ provide better long time solutions, despite the affected repeatability and frequency range. Piezoelectric accelerometers have low internal damping with resonance frequencies around 30 kHz. The usual upper frequency limit for condition monitoring is located between 10-20 kHz and the lowest frequency could be less than 1 Hz [1].

3.3.2 Sampling Frequency

As is widely known in signal processing, sampling is the operation which transform a continuous or analogue signal into a discrete one. The number of samples taken in one second of the signal is called sampling frequency f_S and is expressed in Hz. According to Stamboliska [17], the required sampling frequency is strictly related to the frequency of the signal that needs to be monitored. This rate, should be set in a way which allow the reconstruction of the analogue signal from the discrete data without losing information on the measured parameter.

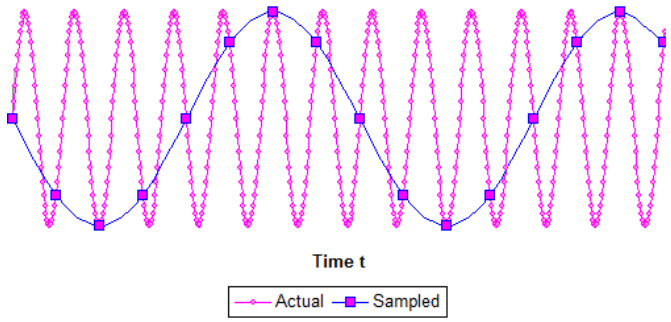


Figure 3.9: Sampling frequency effect and aliasing phenomena

To respect what had been said in the previous paragraph, the Shannon-Kotielnikow condition must be fulfilled, i.e. to make the signal reconstruction without loss of information. This condition states that the sampling rate cannot be less than twice of the highest frequency of the measured analogue signal, or in other words, the Nyquist frequency. When this condition is not fulfilled the phenomenon of aliasing is observed, as exemplified in figure 3.9 obtained by Greg Stanley. Despite the value of the Nyquist frequency, in condition monitoring, the minimum sampling rate must be much higher than the one given by Shannon-Kotielnikov equation. This is also called oversampling and the minimum frequency can be determined based on tests for each practical case. In accordance with the previous author [17], in the stationary signals obtained from low-speed machinery, this frequency is 5-10 times higher than the Nyquist frequency, which in the low-speed equipment (600 rpm or 10 Hz) correspond to a value between 100 and 1000 Hz.

3.3.3 Filtering

A wide band noise is created by running equipments that support or depend on low-speed machinery. This wide band can be generated by electrical motors or fans and cause disturbs on the obtained signal. To minimize their effects, filtering of the measured data is used and for low-speed machinery is common to use low pass filters [17]. It enables to cut higher and not useful frequencies which come from operation of the surrounding equipment. Depending on the data acquisition hardware, the filtering operation can be done in the post-processing operations or online during the measurement. The effect of low-pass filtering is observable on the figure 3.10 which represents a temperature measurement from Satlantic observatory.

To properly design and apply a low pass filter, the most important parameter is the

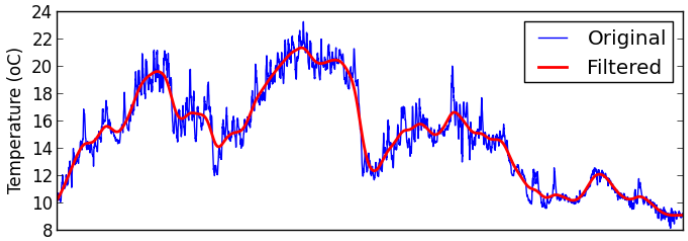


Figure 3.10: Low pass filter applied to a temperature signal

frequency f_{1P}. If this is set too high, the signal will not be properly filtered and high frequency noise will remain. However, if on the other hand, it is set too low, important data can be lost. Theoretically, this frequency is related to the Nyquist frequency, designated f_N, such that the filter frequency is lower than the last. Notwithstanding, considering the efficiency of filters available in the monitoring systems, this frequency is usually lower and the equation (3.7) converges into (3.8), where f_S is the sampling frequency and f_{NP} is the practical Nyquist frequency. In order to achieve the optimum results, the low pass filter can be obtained after conducting several tests with different filters as an iterative process.

$$f_{1P} \leq f_N \tag{3.7}$$

$$f_{1P} \leq\sim 0.5 \cdot f_S = f_{NP} \tag{3.8}$$

3.4 Diagnosis

The generic tools used on CM can be separated into two categories of inspections which are subjective and objective inspections. They are valid for low, medium and high speed machinery. Despite this, the importance and information obtained by them, can vary depending on the machinery speed and principle of operation. The ones discussed in this section suit the low and medium speed cases.

3.4.1 Subjective Inspections

These techniques provide information that use four basic senses which are look, listen, feel and smell. Some of the more experienced maintenance personnel widely use these tools every time they pass by an equipment. These senses play a very important role in the low speed machinery because components moving at low speed allow a good visual inspection and an overview, as well as sound distinction of the standstill parts and moving components [17].

The low speed movements usually produce lower noise that provide better conditions for noticing any irregularities such as dry friction, strokes, knocking and others. Also the visual inspections can detect leakages, cracks, broken or missing items or assess lubrication condition. The feel can be used to detect higher temperatures or vibrations on the bearings housings. The smell allow the detection of burnt items or gasses leakages. The information provided by the basic senses is very useful once it is used as an alarm or trigger for measurements and corrections. However, they cannot be used as a base for a condition monitoring strategy because

those requires more precise data in order to monitor trends and determine the exactly level of certain parameters. This was the motto for developing other measurements and strategies to assess the conditions of the equipment in the past few years [17].

3.4.2 Objective Inspections

The objective inspections, require special instruments and skilled personnel as well as documented systems to gather all the data in an historical way and consequently enable detailed engineering evaluation, such as statistical analysis of the results [17]. There are many sources to conduct these inspections but here it will be discussed the most important ones for rolling element bearings under low and medium speed rotation.

Dynamic Effects - in this technique, monitoring detects potential failures which cause abnormal amounts of energy, emitted in form of waves such as vibration, pulses and acoustic effects from rotational and reciprocating equipments, as well as stationary structures or equipment being under influence of external vibrations [17]. There are several techniques which can be applied as detailed in Chapter 4. However for low speed machinery special considerations must be taken into account about the instruments and estimation rules because the dynamic effects differ from medium and high speed machines.

Particle Effects - in this case, the monitoring detects potential failures which cause discrete particles of different sizes and shapes to be released into the environment in which the item or component is operating [17]. Among others, the most used techniques are analytical ferrography, mesh obscuration particle counter and magnetic chip detectors. They are useful to asses machine degradation, oil contamination or oil deterioration. It is suitable for all ranges of speeds.

Physical Effects - evaluates failures that cause changes in the physical appearance or changes in the structure of the equipment which can be detected directly. The monitoring techniques detect potential failures in the form of deformations, cracks and fractures. It is used for examination of solid bodies of various equipments such as weldings or pipelines The most used techniques are liquid dye penetrants, electrostatic fluorescent penetrant, ultrasonic techniques and X-Ray radiography.

3.4.3 Signal Models

The faults that occur on bearings may be separated in a generic way between localized and extended, because of that, and the way to analyse a faulty bearing signal depends on the type of fault present. The initial small localized faults give rise to sharp impacts and the extended faults make the impacts smoother. Randall [1] present two models for the signal generated by a bearing fault. The first one assuming a truly cyclostationary signal with the uncertainty of occurrence independent of the number of periods into the future. The second one, give a pseudo-cyclostationary signal with the uncertainty of occurrence increasing with the number of periods into the future.

Localized Faults

The main issue with localized fault is how to model the random spacing of the impacts. The first publication to model bearing fault signals as cyclostationary was not very convincing because the main resonances excited by the faults were outside the measured range [10]. However, the way of modelling the random variation in pulse spacing was later found to be incorrect and more suitable with the second model explained before. In the original model, a random component was set around a known mean period while in the correct model it was actually the spacing itself that was the random variable.

This had implications on the prediction uncertainty for the location of a future pulse. In the first model, this was constant and in the second model the variation cause by slip and thus the uncertainty increases with time of prediction into the future. This means that the signals from a localised fault in a bearing are not truly cyclostationary but are better defined as pseudo cyclostationary. The practical consequences of this are explained in figure 3.11. It is observed that in terms of interpreting the spectrum, there is little practical difference in treating the pseudo-cyclostationary signals as cyclostationary and that is how we will proceed afterwards [10].

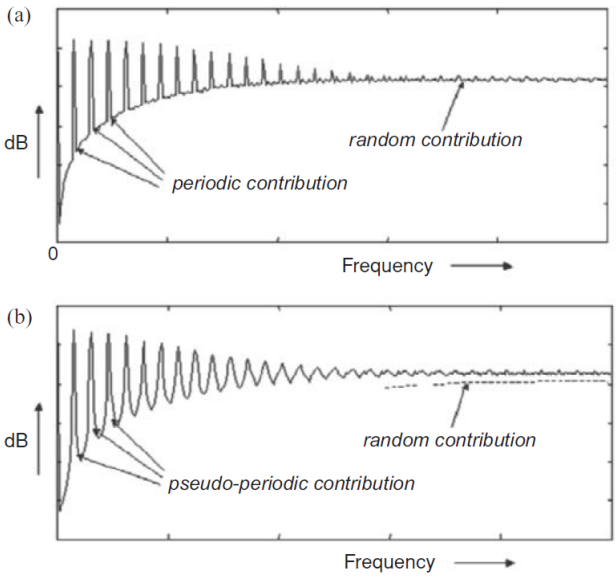


Figure 3.11: Comparison between the two models for localized faults [10]

Extended Faults

For extended spalls, there will often be an impact as each rolling element exits the spall and in that case the envelope analysis, discussed in Chapter 4, will often reveal and diagnose the fault and its location. Despite this fact, there is a tendency for the spalled area to become worn and in this case the impacts might be much smaller than in the early stages. The typical modulating signal, shown in figure 3.12 contains the local mean value and amplitude modulated noise cyclostationary components. This mixture would have a spectral correlation with discrete characteristics in the cyclic frequency direction, but a mixture of discrete and continuous characteristics in the normal frequency direction. However, if the periodic components are removed, only the second order cyclostationary components will be

left and they could only come from an extended bearing fault [10].

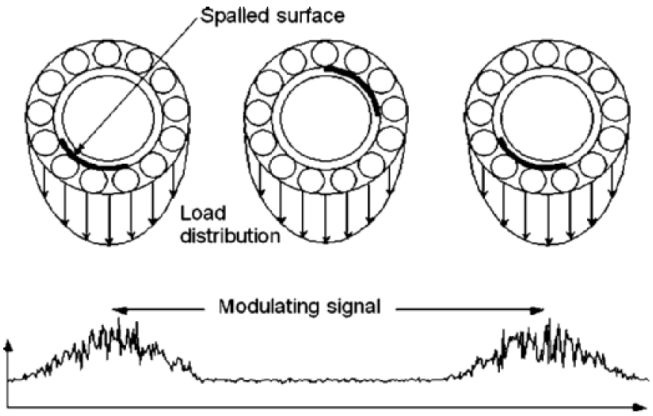


Figure 3.12: Generation of a modulating signal by an extended inner race fault [10]

3.5 Prognosis

Prognosis has been defined, by the International Standard Organization, as the estimation of the time to failure and the risk for one or more existing potential failures to develop into active ones [3]. Unlike the diagnosis, prognosis can be considered as a prior event analysis which aims to predict the residual life of a component before the failure occurs, along with the evolution of wear in the entity. There are two main prediction types in machine prognosis. The most used is the Remaining Useful Life (RUL) which is the time left before observing a failure. This is essentially a prediction of how much time is left before a failure occurs, given the current machine condition and past operation profile [23]. The other is to predict the chance that a machine operates without a fault or a failure up to some future time, which is critical in some situations where a failure would be catastrophic such as a nuclear plant. This section will be addressed the first technique which is the RUL. Prognosis is based upon trending of vibration parameters and as will be seen in Chapter 5, there are a large number of statistical parameters available. The problem is which parameter should be trended for a particular machine class [24].

3.5.1 Wear Evolution

In order to understand the condition evolution of the equipment it is necessary to know the different stages experienced by the components before failure. The one addressed in this section is the rolling element bearing. Initially, back in 1986, there were defined three stages to classify the bearing life. However, this number has increased to four in 1991 [24].

- Stage 1 - in this stage, sometimes refereed as prefailure, the problems appear in ultrasonic frequencies from about 250-350 kHz. With the wear evolution these values decrease and reach 20-60 kHz frequencies. Acquiring high frequency enveloped spectra confirms whether or not the bearing is developing a failure.

- Stage 2 - slight bearing defects begin to bring bearing component natural frequencies to the sight. These stage, situated between the prefailure and the damaged one, can be defined as intermedium. The spectrum is dominated by the defects in the 500-2000 Hz range. Sidebands appear above and below natural frequency peak at the end of this stage. The overall spike energy increase.

- Stage 3 - the bearing defect frequencies and harmonics appear. With the wear progress, more defect frequency harmonics appear and number of sidebands grow, both around these and the bearing natural frequencies. This is called the failure stage and the wear is usually visible and may extend throughout periphery of bearing, particularly when many sidebands accompany bearing defect frequency harmonics.

- Stage 4 - also known as catastrophic failure, it affects the shaft speed components in the spectrum. It grows and normally causes the growth of many running speed harmonics. Discrete bearing defect and natural frequency components, beginning to disappear and being replaced by random and broadband high frequency noise floor.

The analyst may use this information about the bearings stages [17] to place the ones he is trying to evaluate and afterwards estimate the RUL as it will be seen later.

3.5.2 Remaining Useful Life

RUL, also called remaining service life, residual life or remnant life, refers to the time left before observing a failure, given the current machine age and condition, and the past operation profile [23]. It is defined as the conditional random variable expressed by equation (3.9), where T denotes the random variable of time to failure, t is the current age and $Z(t)$ is the past condition profile up to the current time.

$$T - t | T > t, Z(t) \tag{3.9}$$

To correctly interpret the random variable, the distribution of it is essential. The fault propagation process is usually tracked by a trending or forecasting model for certain condition variables [23]. There are two ways in describing the failure mechanism, the first one assumes that failure depends only on the condition variables, which reflect the actual fault level and the predetermined boundary. The proposition in this case states that the failure occurs when the fault reaches a predetermined level. The second way builds a model for the failure mechanism using historical data. In this approach, different definitions of failure can be used, which are an unsatisfactory level where the machine is operating, a functional failure when the machine cannot perform its intended function at all or it can be just a breakdown when the machine stops operating. According to [23], the machine life can be separated into two intervals. The first one called installation to potential failure interval in which the machine is running correctly and the second one, called potential failure to functional failure, in which the machine is running with a problem. Based on the distributions estimated for each interval, it is possible to predict the time to failure or RUL.

3.6 Wind Turbines

The wind energy industry has been growing rapidly in the last 20 years. In fact, the market forecasts that by 2018 the wind energy cumulative gigawatts (GW) will be 43% higher than in 2015 [2]. This renewable energy is a reliable energy source and according to the World Wind Energy Association (WWEA) its application, one of the most promising in the field. The global wind capacity was, back in 2014, around 370 GW with China alone installing over 23 GW [25]. According to the same study, in the state of Iowa in the United States, the electricity generated by wind power was over 28% and in 2013, Spain got 69% of their demand from wind energy. The European Wind Energy Association (EWEA) is planning to expand the offshore capacities for more than 98 GW.

According to the 2015 report of EWEA there has been a decrease in the use of conventional power sources such as fuel oil and coal and an increase of 6.3% on installations of wind turbines compared to the 2014 installed capacity. Germany is the country with more installed capacity 45 GW, Spain is next with 23 GW, followed by the United Kingdom and France with 14 and 10 GW respectively. Portugal takes the 8th place in the table with the same installed capacity as Poland and Denmark of 5.1 GW and after Italy and Sweden with 9 and 5 GW installed respectively. The investment in Europe, in 2015, to finance wind energy development was 26.4 billion € which represented an increase of 40% compared to 2014. This values let us have an idea of the future and the tendency in this energy source [26].

Despite all the investment being made in wind turbines development there are still some improvements to be made in order to make this source of energy competitive in terms of availability, reliability and the expected life of the turbines. With the increasing of the demanding for renewable sources of energy, due to increased public awareness of climate changes or to government policies, the management costs will have to be carefully taken care off. According to Márquez [27], for a 20 year-life, the operations and maintenance costs of 750 kW turbines might account for about 25-30% of the overall energy costs or 75-90% of the investment costs.

For instance, early cracks in bearings have been detected with less than 3 years of operation which corresponds to 5-10% of the expected life [2]. These values are even worse in offshore applications, with the increasing of the turbine size operating on harsher environments. Figure 3.13, presenting information collected between 1993 and 2004, suggest that larger turbines fail more frequently [27] and consequently the maintenance costs increase.

Due to this fact, there is an urge for the engineers to find a way of reducing wind energy costs and make the turbines smarter and able to surpass the working environment challenges such as the rapidly temperature changes, the air pressure and the alternating load. Condition monitoring has been strongly applied to wind turbines to make sure that the components continue to perform the functions for which they were designed. A wind turbine may experience fault in several systems such as the gearbox, the generator, the bearings, the blades, the structure or even the foundation. A scheme of a wind turbine is presented in figure 3.14 where the number from 1 to 7 represent respectively the blades, rotor, gearbox, generator, bearings, yaw system and tower [27].

This text is specific for bearings and it will focus on the bearing failures on the wind turbines, despite other failures may arise. This component is responsible for around 70% of gearbox downtime and 21-70% of generators downtime, 21% in small generators, with a power lower than 1 MW, 70% on medium generators, with a power between 1 and 2 MW, and 50% on large generators with powers greater than 2 MW [2]. Other typical failure is in the mechanical braking system which is generally applied after the blade furling and electromagnetic braking have reduced the shaft speed, in some countries, the wind turbines are required by law to have two independent fail-safe brake mechanisms to stop the turbine promptly when required [28].

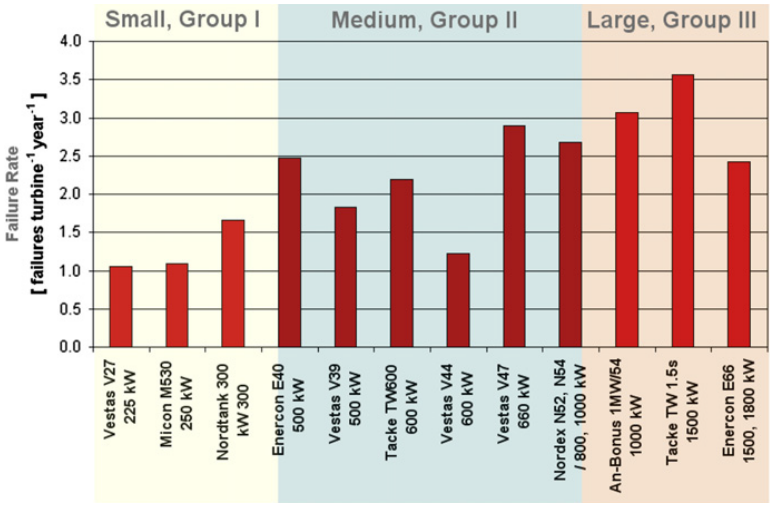


Figure 3.13: Distribution of failure frequencies according to turbine type [27]

Currently the condition monitoring techniques applied to the wind turbines are vibration analysis, acoustic emission, ultrasonic, oil analysis, termography, radiographic inspection or electrical effects. The most popular is the vibration analysis where different sensors are spread along the wind turbine. The type of sensors used are diverse, there are transducers for the low-frequency range, velocity sensors in the middle frequency and accelerometer for the high frequency range. Other well spread technique is the oil analysis which for ultimate purpose, stablish the oil quality by taking periodic lubricant analysis. Despite the advantages and disadvantages of each technique, they all complement together [27]. Actually, the most common practice in wind turbine condition monitoring is the Supervisory Control and Data Acquisition (SCADA). This approach, monitors the conditions of the wind turbine based on their historical data, including wind speed, temperature and power outputs [29]. Some SCADA systems include vibration and oil analysis in their plans in order to provide more reliable data. In other words, the SCADA system relates the wind turbine operating conditions with the condition of the system and measure the rate at which the damage is occurring [29].

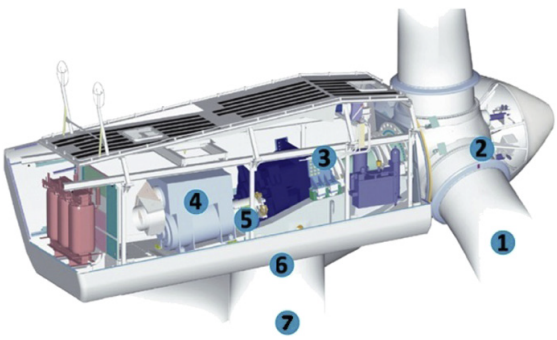


Figure 3.14: Main parts of a turbine [27]

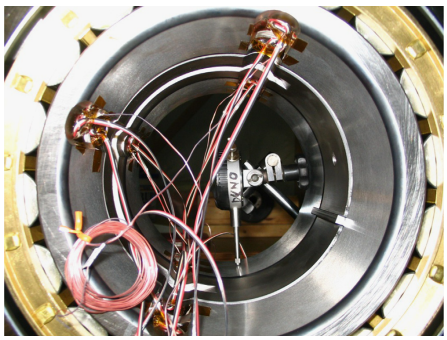


Figure 3.15: Instrumentation of a gearbox bearing [27]

3.6.1 Technical and Financial Challenges

The latest advantages in signal processing have improved the condition monitoring technologies on the diagnosis and prognosis of wind turbines while reducing the operational costs. However, there is still a long way to go and several improvements to be made. The first of those improvements is in the purchasing phase, due to the fact that the payback of the system cannot be exactly measured before application and fault detection. In addition to that, some wind-farms managers are also discouraged by the complexity of the systems. However, the price and financial gains can be calculated based on data and probabilities from other wind-farms.

On the topic of monitoring techniques there is a need to become more flexible, the integration of different methods in one system. Currently there is a drawback in the interpretation of datasets and results arising from different measures. While an automatic system is not available there will still be false alarms and misdiagnoses in condition monitoring strategies. The automatization will also allow the application of a complex strategy to a large number of turbines, in other words, became the system scalable so that several wind turbines can be monitored simultaneously. In wind turbines, due to the large components and tools involved, there is a need to maintain a tight maintenance schedule, so the need to early prediction gets more relevant. The logistics involved can take a reasonable amount of time and the interval between detection and failure may not be enough.

Setting alarm levels and the optimum monitoring approaches, capable of performing isolated monitoring without previous information, is a crucial need, since currently there is the requirement to have historical data. For last, predicting the remaining useful life for the components with a great accuracy would allow a more controlled maintenance and reduce the costs even more, because the components could be running with lower security factors.

Chapter 4

Data Processing Techniques

4.1 Basic Techniques

4.1.1 Time Domain

The representation of the data acquired during the measurement as a plot of amplitude versus time is referred as a time-domain profile [30]. Being one of the simplest and easiest ways to represent data and perform diagnosis, this approach is overall used for linear and reciprocating machinery. Despite that, for no experienced users, this approach can be difficult to use due to the fact that the vibration data in this domain is added together to represent the total quantity of the parameter being measured. This makes it harder to directly observe the contribution of a single vibration source. The applicability range of this technique may be as simple as visually observing the signal to more refined approaches such as trending time domain statistical parameters. The last was the main focus of this text and will be properly discussed in the Chapter 5.

If properly understood, time domain analysis can retrieve enormous amounts of information. Figure 4.1 shows the type of analysis that can be conducted on time domain, the probability density moments will be discussed after and also superficial clarification on the others will be conducted.

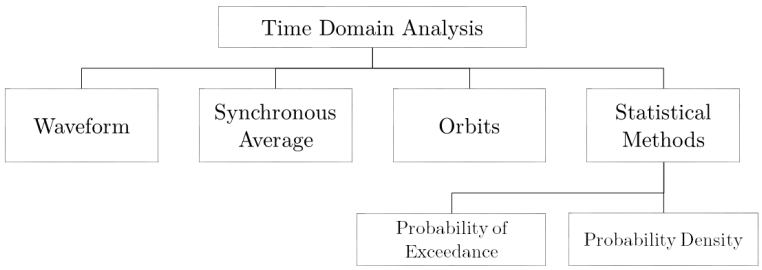


Figure 4.1: Techniques used for time domain analysis

Waveform

The simplest technique of visually looking for portions of the time waveform, allows the user to determine the presence of amplitude modulation, shaft components, unbalance, transients or even higher frequency components [24]. Waveform analysis is based on the raw signal directly collected from the transducer. In vibration terms, it is the representation of displacement, velocity or acceleration with time.

The most common use of time waveform is to compare the pattern of one equipment with another, obtained from a machine with similar defects. One phenomena widely identified on the time domain is the beat [4] that is represented in figure 4.2. This occurs when two waveforms are separated only for slightly frequencies, which can be lower than 30 Hz, and have the similar amplitude. These are pulses due to alternating reinforcement and cancellation of amplitudes. The change in amplitude is called modulation and it is characterized by a frequency corresponding to the difference between the frequencies of the two waveforms [11]. The beat phenomena is common at centrifuge machinery which has plates running at different speeds and is common to observe the beat if any unbalance is present. Another example are the electrical motors with defects in the electrical part. In electrical motors, the second harmonic is very notorious [11] which means that if the power supply frequency is 50 Hz the second harmonic will be 100 Hz or 6000 rpm, knowing that the physical speed of the motor will never achieve the maximum provided by the grid. Lets assume that the engine is running at 2980 rpm which will have a second harmonic of 5960 rpm, due to the difference of this and the 6000 rpm component, the beat will be present.

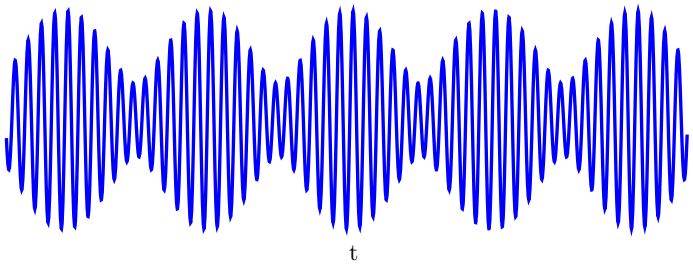


Figure 4.2: Example of the beat phenomena

There are some fields where the time domain analysis provide better additional information than the one obtained with the frequency analysis. Low speed applications, assessment of the severity of defects in rolling element bearings and gears, looseness and beats are some of them. However, when a time waveform analysis is carried out, there are a number of parameters that should be taken into consideration. Those are, the unit of measurement, the time of the measurement, the resolution and the time window analysed [11]. According to the same author the data collected should correspond to at least 5-10 cycles of the machine under analysis.

Synchronous Averaging

Synchronous time averaging corresponds to an analytic technique used to locate sources which may be responsible for vibrations in the structure. It is performed by collecting time waveforms with a trigger to one shaft revolution. In other words, is the time signal averaged over a large number of cycles and synchronous with the running speed of the machine [4]. This technique is used to measure vibrations directly and harmonically related to the turning speed

of a specific component or shaft and is also used to remove the background noise as well as periodic events not exactly synchronous with the equipment in focus [11].

Time synchronous averaging is often used when there are more than one shaft present at the equipment being monitored, running at different speeds, which allows the user to remove the non-synchronous components from the signal and only save the vibration from the shaft of interest. This is the reason why a typical application of this method is for gear fault diagnosis where multiple shafts are present [4].

The technique uses a tachometer or other sensor capable of trigger the acquisition system to start measuring having the shaft rotation as reference. As had been said, this technique measures all the vibration synchronous with the reference shaft and that is one of the reasons why it is used to removed those components and its harmonics, which will improve the ability to detect faults on bearings. In this way, the method is used to removed part of the signal and the discrete components removal can be compared to what is done on the cepstrum editing. Other non-synchronous faults such as cavitation in turbines and electrical noise can also be easier detected using the synchronous averaging to clear the signal. On the other hand, for synchronous problems such as unbalance, misalignment or rotor looseness, the method can be applied to extract the components related to the shaft speed and therefore the vibration recorded can be directly observed on the spectrum. As a result of being an averaging technique, a number of averages is required to remove non-synchronous peaks and allow the peaks to average out of the spectrum. According to Scheffer [11] a number between 100-1000 averages are needed to properly perform this technique.

Orbits

The orbits, also known as Lissajous patterns, correspond to the plots of two time signals, simultaneously in the x and y axis. The unit of measurement is displacement which is measured using proximity probes and the pattern obtained is the shaft orbit which can be used to detect bearing wear, misalignment, unbalance and instabilities. This method is particularly useful in turbomachiney to properly balance the shaft which rotate at very high speeds.

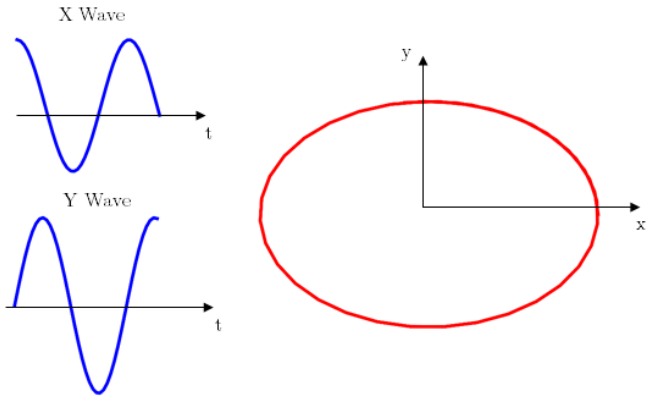


Figure 4.3: Orbits formation with two time signals

Measurements done with the proximity probes are relative because they are considered vibration measurements of the shaft with respect to the bearing housing [11]. This means that the orbit achieved give a visual representation of the shaft centreline movement inside the bearing housing. Acceleration and velocity transducers can also be used, those are referred as external transducers and require mounting on the outside of the bearing housing and produce *case orbits*. This allows absolute shaft motion, relative to space. Figure 4.3 pretend to clarify

the way how orbits are obtained. Considering the two signals on the left, acquired from the transducers placed in the x and y axis, represented in the time domain, the orbit on the right is obtained by plotting the two directions against each other.

Statistical Methods

Time domain statistical parameters have been used to trending analysis in order to detect the presence of incipient bearing damages. There are currently a considerable number of features that can be extracted from the signal and detailed analysis will be focused on the Chapter 5. In this section, an introduction to the topic will be made. The way in which the instantaneous value of the signal is distributed can be expressed in terms of the probability distribution or the probability such that their value is less or equal to a specific value [1].

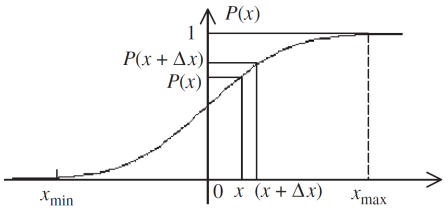


Figure 4.4: Probability distribution for a random signal [1]

Considering a random signal $x(t)$, with a minimum value x_{Min} and a maximum value x_{Max} sampled at small uniform intervals Δt, the probability distribution function $P(x)$ is defined as the fraction of these samples that are less than or equal to a particular value of x. If now we extend this approach to continuous functions, it means that for a time function, $P(x)$ can be interpreted as the fraction of time that $x(t)$ is less or equal to x as perceptible in equation (4.1).

$$P(x) = P\ [x(t) \leq x] \tag{4.1}$$

$P(x)$ must be in conformity with the figure 4.4 which states that $x(t)$ is certain to be less than or equal to the maximum value x_{Max} or in other words $P(x_{Max}) = 1$ and may never be less than the minimum value x_{Min} which is the same as $P(x_{Min}) = 0$. The probability that $x(t)$ is between x and $x + \Delta x$ is given by $P(x + \Delta x) - P(x)$ [1]. The probability density $p(x)$, which corresponds to the slope in the probability distribution curve, can therefore be defined as presented in equation (4.2).

$$p(x) = \lim_{\Delta x \to 0} \frac{P(x + \Delta x) - P(x)}{\Delta x} = \frac{dP(x)}{dx} \tag{4.2}$$

When the calculation of the total area under the probability density curve is performed, the result must always be one.

$$\int_{-\infty}^{\infty} p(x)dx = \int_{-\infty}^{\infty} \frac{dP(x)}{dx}dx = P(x)\Big|_{-\infty}^{\infty} = P(\infty) - P(-\infty) \tag{4.3}$$

As had been said $P(\infty) = 1$ and $P(-\infty) = 0$ so the equation (4.3) returns the value 1 for the area under the probability density curve. The probability of exceedance, also used in condition monitoring of bearings, is characterized as the integral of the probability density curve and gives the probability that the instantaneous vibration amplitude exceeds any particular amplitude [4]. For signals defined as Gaussian random signals, which have a normal distribution, the probability density function is given by the equation (4.4) which corresponds to an exponential curve $\exp(-x^2)$ centred on the mean value and scaled in the x and y direction in terms of the standard deviation σ in order to make the total area equal to the unity.

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{4.4}$$

4.1.2 Frequency Domain

The vibration characteristics of any kind of machinery are on a certain way unique. This is due to the fact that there are different transfer characteristics of each machine that make it different from each other, even if geometrically identical. The method of assembly, mounting and installation of the equipment play a crucial role in its vibration response [4]. It is common practice to collect the signature spectrum when the equipment is running under normal conditions in order to provide the basis for later comparisons and consequently detect significant increases in the vibration levels.

From a practical point of view, the simple harmonic vibration functions are related to the frequencies originated from the rotating components. This means that those frequencies are some multiple of the basic running speed of the machine shaft. Determine these frequencies is the first step in analysing the operating condition of the equipment [30]. It is usual to refer to the power spectrum in vibration analysis on the frequency domain. If a discrete time signal $x(t)$ represents a sampled periodic function with period T it is possible to represent the function by a series of sinusoids as studied by Jacques Fourier, who was a French mathematician. The Fourier series expansion of the function $x(t)$, which is given by equation (4.5), can be obtained from the Fourier integral, where the coefficients a_n and b_n are given by equation (4.6) and (4.7) respectively.

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t) \tag{4.5}$$

$$a_n = \frac{2}{T_0}\int_t^{t+T_0} x(t)\cos(n\omega_0 t)dt \qquad n = 0, 1, 2, ... \tag{4.6}$$

$$b_n = \frac{2}{T_0}\int_t^{t+T_0} x(t)\sin(n\omega_0 t)dt \qquad n = 0, 1, 2, ... \tag{4.7}$$

And ω_0 represents the fundamental angular frequency of a periodic time history and is given by $\omega_0 = 2\pi/T_0$. After calculating all the coefficients, the signal can be represented in the frequency domain as given by equation (4.5). The amplitude of the function at it nth harmonic

can be determined by equation (4.8). The Fourier series of a function, can also be represented in the complex domain and in either the real-imaginary or magnitude-phase format [7].

$$A_n = \sqrt{{a_n}^2 + {b_n}^2} \tag{4.8}$$

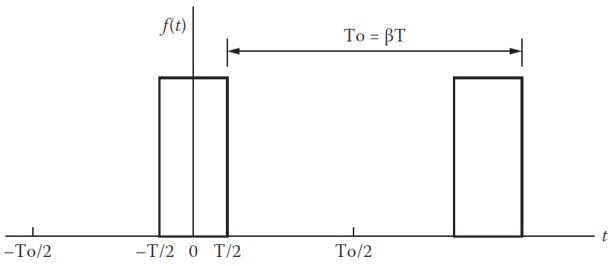


Figure 4.5: Pulses separated in time [7]

The signals in real world applications are not periodic at all times as had already been mentioned. The Fourier coefficients of two pulses separated by a time $T_0 = \beta T$ as shown in figure 4.5 are given by equations (4.9) and (4.10).

$$a_n = \frac{2A}{n\pi} \sin \frac{n\pi}{\beta} \qquad n = 1, 2, 3, ... \tag{4.9}$$

$$b_n = 0 \qquad n = 1, 2, 3, ... \tag{4.10}$$

As the time interval T_0 increases in comparison to the pulse duration, the amplitude of the Fourier coefficients decrease. If the value of T_0 tend to infinity it is possible to obtain the Fourier integral expression of any function that is not periodic, also known as the Fourier Transform (FT). Defining a new coefficient as $A(\omega) = a_n T_0$ and substituting for T_0 tending to infinity, $\omega_0 \approx d\omega$ and $n\omega_0 \approx \omega$ we have the expression.

$$A(\omega) = \lim_{T_0 \to \infty} a_n T_0 = 2 \int_{-T/2}^{T/2} f(t) \cos(\omega t) dt \tag{4.11}$$

For $T \to \infty$ the result of $A(\omega)$ and $B(\omega)$ are given by equations (4.12) and (4.13) respectively.

$$A(\omega) = 2 \int_{-\infty}^{\infty} f(t) \cos(\omega t) dt \tag{4.12}$$

$$B(\omega) = 2 \int_{-\infty}^{\infty} f(t) \sin(\omega t) dt \tag{4.13}$$

Substituting in equation (4.5), one obtains the Fourier integral expression of any function that is not periodic which is given by equation (4.14) and the respective amplitude is given by equation (4.15).

$$f(t) = \frac{1}{2\pi} \int_0^\infty A(\omega)\cos(\omega t) + B(\omega)\sin(\omega t)\ d\omega \tag{4.14}$$

$$F(\omega) = \sqrt{A(\omega)^2 + B(\omega)^2} \tag{4.15}$$

Digital fast Fourier analysis have become the most popular method to obtain the frequency domain signal and will be addressed forward in this text. One of the main purposes in computing the spectrum of a vibration signal is to identify the major frequency components and use these components for trending purposes. For bearing fault detection, it requires the previous knowledge of the FCF which were detailed in the Chapter 2. However, those frequencies are used for localised race damage and can not detect other faults such as lubrication starvation, contamination or misalignment.

Using spectrum comparison, or in other words, plotting the signature spectrum obtained when the machine was free of faults against the faulty spectrum, it is possible to observe the difference between the baseline and the subsequent spectrum [24]. For bearings, the changes are often more notorious in the higher frequency ranges due to the increases in the amplitudes of the structural resonances. Howard [24] suggests that increases of 6-8 dB should be considered as significant and a change of 20 dB in the spectrum magnitude should be considered serious. The same author further states that the comparison should be performed on a logarithmic scale. This representation is also an added value if performed on both axis due to the small changes in shaft speeds that can shift the position of the peaks at different spectrum lines. One way of representing the successive spectrum obtained at different stages of the machine condition is in a 3-axis plot with time, frequency and magnitude. This representation may allow a fault to be more easily recognized. An example of this type of representation is the waterfall plot represented in figure 4.6 which represents the evolution of the spectrum of a gearbox, conducted by Davies [4], were the gears were subjected to an overload condition.

Frequency Parameters

In the Chapter 5 it is proposed the extraction of time-domain features to analyse the trends in the machine under varying operational conditions, with the exception of the magnitude evaluation for the harmonics, which due to the fact that is extracted from the frequency domain, may be considered a frequency parameter. In fact, frequency domain analysis is an approach that can reveal additional information that may not be found in the time domain [31]. Although not being studied in the experimental results, these parameters are of enormous interest for future development and will be now detailed.

The frequency domain features are calculated on the basis of a Fast Fourier Transform (FFT) from the time domain. Here we will present two parameters which are the Matched Filter Root Mean Square (MFRMS) and the RMS of the Spectral Difference (RDO).

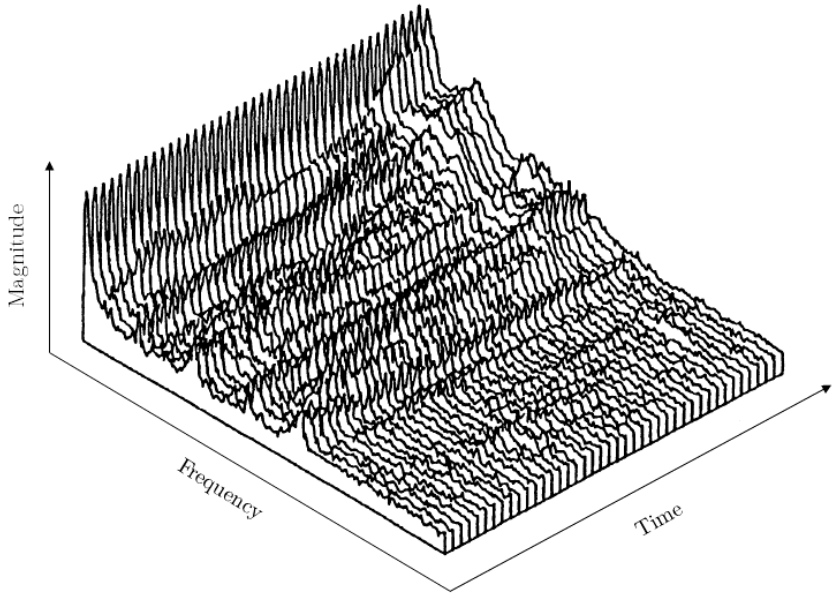


Figure 4.6: Spectrum evolution on a gearbox [4]

1. **MFRMS** - represents a direct comparison of spectral responses where R_n is the amplitude of the n-th line in the reference spectrum an A_n is the amplitude of the n-th line in the current spectrum [24].

$$MFRMS = 10 \log \left[\frac{1}{N} \sum_{n=1}^{N} \left(\frac{A_n}{R_n} \right)^2 \right] \tag{4.16}$$

2. **RDO** - this feature involves finding the RMS of the difference between two spectrum when the amplitude of the spectral responses are given logarithmically [24]. L_n is the level, in dB, of the current spectrum n-th line, while S_n is the level, also in dB, of the n-th line of the reference spectrum.

$$RDO = \left[\frac{1}{N} \sum_{n=1}^{N} (L_n - S_n)^2 \right]^{1/2} \tag{4.17}$$

In some cases, the differences observed between the frequency and time domain parameters can be surprising [4]. However, the accurate analysis of the spectrum may be very complex, due to the defect related harmonics and sidebands present. In addition to that, deciding the frequency bands to evaluate require additional time and can be complicated to rotating equipments [24].

4.1.3 Cepstrum

The cepstrum of a signal was primarily studied in the 1960's as a technique for processing seismic signals in order to separate the influence of echoes from the data. Back then, this was a better alternative than the autocorrelation function for detecting the echo delays. The principal use of the cepstrum for bearing fault detection is to remove the periodic components in the frequency spectrum corresponding to shaft frequency as well as its harmonics and their respective sidebands [24]. In relative terms the cepstrum spots the spectrum periodicities, while this spots the periodic components in the time domain.

The analysis in the quefrency domain is relevant in condition monitoring due to many reasons. The first rahmonic in the cepstrum, indicates directly the harmonic spacing independent of where the harmonics appear in the spectrum. Then, the unknown frequency spacing can usually be traced back to other components such as bearings or gears [24]. Lately, the first rahmonic has been found to be a stable trend parameter, despite not being subject of study in this text. The derivation of the cepstrum is not simple and require a considerable computational effort and expertise in the results interpretation [24]. The common aspect between rolling element bearings and gears is that there is an harmonic series for each characteristic frequencies. When the families of sidebands become many and the signals have poor signal-to-noise ratio, the quefrency analysis is used to recover the signal in the time domain [7]. The spectrum from a rotating machine may be complex, containing several sets of harmonics and sidebands due to frequency modulations, especially due to torsional vibrations. Because the cepstrum has peaks corresponding to the spacing of the harmonics, those can be easily identified, thereon significant peaks in the cepstrum correspond to possible fundamental bearing frequencies [11]. Figure 4.7 represents an example for a case of a steam turbine with missing blades [1]. The study intends to represent the loss of one or a small number of small blades on a large turbine, with little effect on the overall vibration level. Sometimes, the loss of a small blade on a large turbine may cause the opposite blade to lose the material deposits accumulated which may return an approximate balance. This cause several interactions with the stator blades which cause an increase in the harmonics of the shaft speed. In this case, the cepstrum was used to detect and remove the echoes that were overlapping the signal.

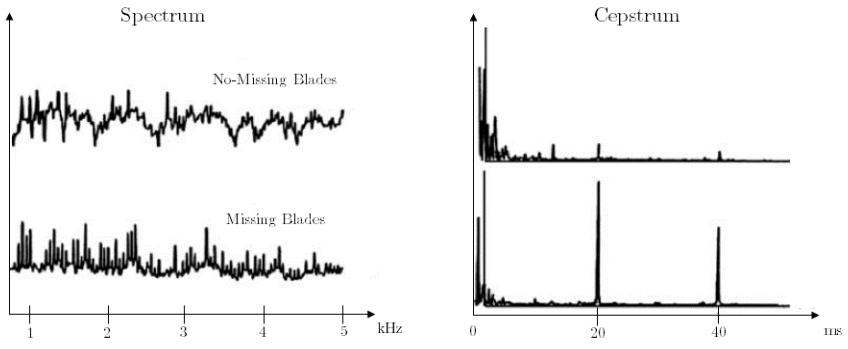


Figure 4.7: Cepstrum used to detect missing blades in a turbine [1]

There are several considerations that must be taken into account when dealing with the cepstrum. The noise level in the spectrum affects the detection of the series of harmonic components. If the harmonics are smeared by operational noise, they will not be detected by the cepstrum. Although a number of techniques can be used to enhance discrete frequency

components, some insurances must be taken during measurements [1]. Another issue arise, when different components are closely together in the spectrum which may cause the higher harmonics to cross over. This may be due to the filter characteristics or the components not being properly separated in the spectrum. One way to solve this would be improve the spectral lines in the data collection.

Another important point is the choice of the vibration parameter. The most common one are velocity or acceleration, with displacement not being commonly used [1]. The effect of the choice between the first two is notorious in the low quefrency because the difference between them is generally a slight change in the lope of the log spectrum, due to the integration from acceleration to velocity [1].

The cepstrum was initially defined as the power spectrum of the logarithm of the power spectrum. Later, it was redefined as the inverse FT of the log power spectrum, because it is more logical to use the inverse transform between a function of frequency and a function of time which makes it reversible to the power spectrum [1]. The complex cepstrum was defined as the inverse FT of the complex algorithm of the complex spectrum. The difference between the autocorrelation function and the cepstrum is the fact that the last is the logarithmic conversion of the spectrum before the second transform. The word cepstrum is originated by reversing the first syllable of spectrum, which was justified because the cepstrum is the *spectrum of a spectrum*. Some examples of the terminology are given in the Table 4.1.

Table 4.1: Terminology in the Quefrency domain

Frequency	**Quefrency**
Harmonic	Rahmonic
Filter	Lifter
Low-Pass Filter	Short-Pass Lifter
High-Pass Filter	Long-Pass Lifter
Magnitude	Gamnitude
Phase	Saphe

As had been said, the cepstrum is defined as the inverse FT of the log power spectrum and is given by equation (4.18).

$$C(\tau) \;=\; \mathcal{F}^{-1}\left[\log(X(f))\right] \tag{4.18}$$

$$X(f) = \;\mathcal{F}\left[x(t)\right] \;= A(f)\; e^{j\phi(f)} \tag{4.19}$$

In terms of amplitude and phase,

$$\log X(f) = \ln A(f) + j\phi(f) \tag{4.20}$$

When the $X(f)$ is complex, the cepstrum provided by equation (4.18) is known as the complex spectrum. When the power spectrum is used to replace the spectrum $X(f)$ the resulting cepstrum is known as the power cepstrum or real cepstrum and is given by equation (4.21).

$$C_{xx}(\tau) \;=\; \mathcal{F}^{-1}\left[2\ln A(f)\right] \tag{4.21}$$

Before calculating the complex spectrum, the phase function $\phi(f)$ must be unwrapped to a continuous function of frequency, what can be complex, making the real cepstrum more easy to use. Another type of cepstrum which is used is the differential cepstrum, which is defined as the inverse transform of the derivative of the logarithm of the spectrum [1]. It is calculated using the z-transform and is given by the equation (4.22) where n is the quefrency index.

$$C_D(n) = Z^{-1}\left[z\;\frac{\frac{d}{dz}H(z)}{H(z)}\right] \tag{4.22}$$

Or as directly calculated from the time signal

$$C_D(n) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}(n\;x_n)}{\mathcal{F}(x_n)}\right] \tag{4.23}$$

4.1.4 Deterministic-Random Signal Separation

The signal coming from a machine consists of several origins and that is the reason why a considerable part of condition monitoring deals with separating the components providing from individual sources. One signal can be defined as a mix of deterministic, for example from gears, and random signals. The random signals may be background noise on the sensors, sources such as the fluid flow in a pump or turbine or cyclostationary signals, as the case of rolling element bearings, which are the main focus of this work. Even with machines running at constant speed, the signals connected to the shaft may not be truly deterministic, unless random speed variations are removed.

4.1.4.1 Angular Resampling

In cases where there are variations in the shaft speed, a technique called order tracking allows the sidebands on the harmonics from the speed to be detected. If a signal, which is synchronous with some phenomena, for instance the rotation of the shaft, is sampled at a fixed number of times per revolution, the digital samples are not separable from those of a sinusoid, which in other words means a line in the spectrum. On the other hand, if normal temporal sampling is used, the spectrum spreads over a range, corresponding to the variation in the shaft speed and consequently a broadband. To perform this method, it is necessary to generate a sampling signal from a tacho or shaft encoder which is synchronous with the phenomena of interest, in this example, the shaft speed. However, this approach has a finite response in time and may not measure the slight random speed fluctuations along the cycle [1].

Due to the limitations state in the previous paragraph, it is best to resample each record digitally based on the respective period of the tacho signal, this allows sampling for uniform increments in shaft rotation angle, also called angular resampling. This technique may be used in several ways, based on digital interpolation. The simplest way is to increase the sample rate

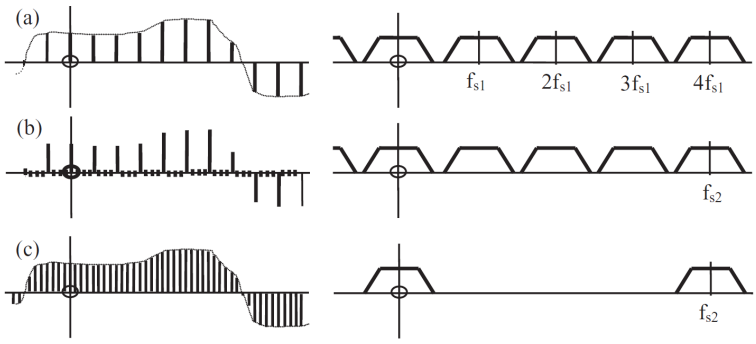


Figure 4.8: Digital resampling with four time the original sampling frequency [1]

by a large factor, for example 10, and after that select the nearest sample to each theoretical interpolated position. This can be done in the time domain, by inserting the needed number of zeros in between each actual sample and then applying a low-pass filter to limit the frequency range to the original maximum, and smoothing the curve. This method is illustrated by figure 4.8 where in (a) is observable the original signal, sampled at frequency f_{S1}, the addition of zeros to change the sampling frequency in (b) and the final result after low-pass filtration and rescaling in (c).

Another way to achieve this is in the frequency domain, by filling the spectrum with zeros in the centre and then inverse transforming the increased spectrum to the same increased number of time samples. A better interpolation, not limited to a ratio of integer numbers, can be achieved by fitting a curve to a group of samples and then calculating the value of the polynomial at the interpolated positions. The accuracy of this approach may be assessed by considering that the interpolation in the time domain corresponds to a multiplication in the frequency domain by a filter characteristic, which aliases back into the measurement range. Another way to perform angular resampling is to use phase demodulation of some synchronous phenomena related to the one of interest, in order to obtain a mapping o shaft rotation angle versus time. If order tracking is being performed directly on an analogue signal, it must be ensured that the signal is previously low-pass filtered to prevent aliasing, in particular when resampling at a lower frequency [1]. In addition to this, digital filtering may also be used as the cutoff frequency varying with the sampling frequency, never forgetting the initial low-pass filter to avoid the aliasing components of entering into the measurement range.

It is illustrated, in figure 4.9 the effect of order tracking on the spectrum of a signal from a gearbox, with a notorious variation in speed over the operational cycle. Without order tracking, no discrete frequency components were visible in the spectrum. This study was conducted in a mining shovel by Robert Randall.

4.1.4.2 Linear Prediction

Another way to obtain the deterministic part of the signal is based on a certain number of samples in the immediate past, and then using that to predict the next value in the series. The random part of the signal, or residual, is consequently obtained by the subtraction from the actual signal value [1]. The model used for linear prediction is an autoregressive model described by equation (4.24).

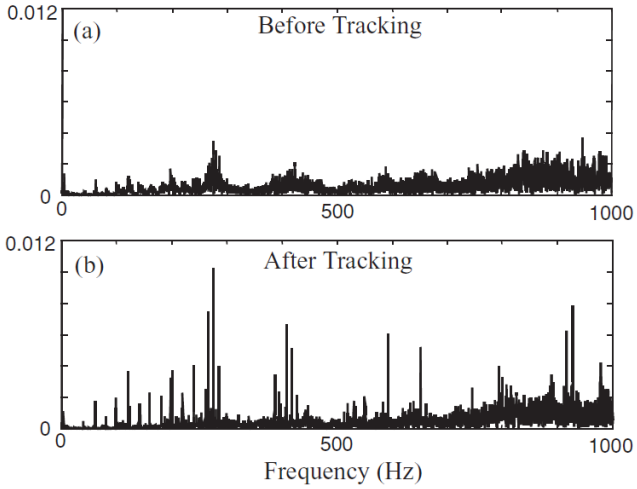


Figure 4.9: Example of order tracking to reduce speed variations influence [1]

$$\hat{x}(n) = -\sum_{k=1}^{p} a(k)x(n-k) \tag{4.24}$$

Where the predicted current value $\hat{x}(n)$ is obtained as a weighted sum of the p previous values. The actual current value is therefore given by the sum of the predicted value with a noise term $e(n)$.

$$x(n) = \hat{x}(n) + e(n) \tag{4.25}$$

The $a(k)$ weighting coefficient can be obtained by a linear operation from the autocorrelation function R_{xx}, stated in the Chapter 2, of the time series $x(n)$, for which trending estimates can be obtained from equation (4.26).

$$\hat{r}_{xx}[k] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]\ x\ [n-k] \qquad 0 \leq k \leq p-1 \tag{4.26}$$

The $a(k)$ are obtained using the Yule-Walker equations, the matrix given by equation (4.27), usually applying the Levinson-Durban recursion algorithm, or also using the Burg maximum entropy method [1].

$$\begin{bmatrix} r_{xx}[0] & r_{xx}[-1] & \dots & r_{xx}[-p+1] \\ r_{xx}[1] & r_{xx}[0] & \dots & r_{xx}[-p+2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[p-1] & r_{xx}[p-2] & \dots & r_{xx}[0] \end{bmatrix} \begin{bmatrix} a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix} = - \begin{bmatrix} r_{xx}[1] \\ r_{xx}[2] \\ \vdots \\ r_{xx}[p] \end{bmatrix} \tag{4.27}$$

The equations (4.24) and (4.25) can be combined and rewritten as stated in equation (4.28).

$$x(n) + \sum_{k=1}^{p} a(k)\ x\ (n-k)\ =\ e(n) \tag{4.28}$$

This can be considered, in the z-domain, as the output $X(z)$ of a system with a transfer function $1/A(z)$ when excited by the forcing function $E(z)$, as stated by equation (4.29)

$$X(z) = \frac{1}{A(z)}\ E(z) \tag{4.29}$$

An important consideration in applying this technique is the choice of the model order p. According to Randall [1], when the application is to separate discrete frequency components from stationary white noise, a standard approach is to use the Akaike information criterion.

4.1.4.3 Discrete-Random Separation

Adaptive noise cancellation is a procedure where a primary signal, containing two uncorrelated components can be separated into those components, using only one of them, the reference one. The reference signal does not have to be similar to the corresponding part of the primary signal, however it has to be related to it by a linear transfer function. This procedure, finds iteratively the transfer function, and can therefore subtract the modified reference signal from the primary one, leaving the remaining component. The self-adaptive noise cancellation is used when one of the two components to be separated is deterministic, and the other random. This enables the reference signal to be assumed as a delayed version of the primary signal [1].

In contrast to the previous approach, the discrete-random separation does not require adaptation, because the required filter to remove the discrete frequency components is previously determined [1], if possible from the whole length of the data to which it will be applied. The basic principle of this technique is to obtain the transfer function between the signal and a delayed version of itself. Similarly to modal analysis, to obtain the H_1, the process uses the cross-spectrum from input to output and divide it by the autospectrum of the input as stated by equation (4.30), where $G_a(f)$ is the spectrum of the input, $G_b(f)$ is the spectrum of the output, $G_{ab}(f)$ is the cross-spectrum and $G_{aa}(f)$ is the input autospectrum.

$$H_1(f) = \frac{E[G_b(f)G_a^*(f)]}{E[G_a(f)G_a^*(f)]} = \frac{G_{ab}(f)}{G_{aa}(f)} \tag{4.30}$$

In perfect conditions, this would return a value of one at the frequencies of the discrete components, due to the correlation of the signals, and zero at noisy frequencies. However, in real life applications, the value depends on the Signal-to-Noise Ratio (SNR). Randall [1], showed that the amplitude of the separation filter is given by equation (4.31), where ρ is the SNR, N is the transform size and $W(f)$ is the FT of the window used, scaled to a maximum value of one in the frequency domain.

$$\frac{(\rho N/2)|W(f)|^2}{(\rho N/2)|W(f)|^2+1} \tag{4.31}$$

4.1.5 Fourier Transfrom

A vibration response can be defined by displacement, velocity or acceleration, and these quantities can be represented in both time and frequency domains. In the frequency domain, the waves have a magnitude and phase variable with the frequency. All the measured vibrations are always in analogue form, or simply, in the time domain and need to be transformed to the frequency domain [11]. This is accomplished using a Fourier Transform to convert the signal $x(t)$ on variable time to $x(\omega)$ on the frequency.

4.1.5.1 Discrete Fourier Transform

The Fourier integral given by equation (4.14), lack the definition of the function $f(t)$ to assess the Fourier coefficients, given by equations (4.12) and (4.13). The sampled time signals are recorded with a finite length, which leads to a situation where the spectrum is discrete and the time record implicitly periodic. Therefore, the infinite integrals of the Fourier integral become finite sums. The Fourier integral can be numerically estimated where the coefficients are obtained at frequency intervals, as expressed by equation (4.32), where the angular frequency ω_n in rad/s has been replaced by the continuous frequency f expressed in Hz while k is an index. N is the length of the record.

$$X(kf) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)e^{\frac{-j2\pi kfn}{N}} \tag{4.32}$$

Dividing by this parameter, the Fourier series components are correctly scaled [1]. This operation can be understood as a matrix multiplication, where X_{kf} represent the vector of N frequency components, while x_n represents the N time samples $x(n)$. W_{kfn} represents a square matrix of unit vector $\exp(-j2\pi kfn/N)$ with angular orientation, depending the rows on the frequency index k and the columns on the time sample index n.

$$X_{kf} = \frac{1}{N}W_{kfn}x_n \tag{4.33}$$

$$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \end{bmatrix} = \frac{1}{8}\begin{bmatrix} \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \uparrow & \nearrow & \rightarrow & \searrow & \downarrow & \swarrow & \leftarrow & \nwarrow \\ \uparrow & \rightarrow & \downarrow & \leftarrow & \uparrow & \rightarrow & \downarrow & \leftarrow \\ \uparrow & \searrow & \leftarrow & \nearrow & \downarrow & \nwarrow & \rightarrow & \swarrow \\ \uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \downarrow \\ \uparrow & \swarrow & \rightarrow & \nwarrow & \downarrow & \nearrow & \leftarrow & \searrow \\ \uparrow & \leftarrow & \downarrow & \rightarrow & \uparrow & \leftarrow & \downarrow & \rightarrow \\ \uparrow & \nwarrow & \leftarrow & \swarrow & \downarrow & \searrow & \rightarrow & \nearrow \end{bmatrix}\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{bmatrix} \tag{4.34}$$

For $k = 0$ the zero frequency value $X(0)$ is simply the mean value of the time samples $x(n)$. For $k = 1$ the unit vector rotate $-1/N$-th of revolution for each time sample increment, which translates in a complete clockwise revolution after N samples. For higher values of k the rotation speed is proportionally higher. For $k = N/2$, also known as Nyquist frequency, or half the sampling frequency, the vector turns $-\pi$ for each time sample, however, it is not possible to assess in which direction it is turning. For $k > N/2$, the vector turns more than π in the negative direction, although it is easier to interpret this as turned in the opposite direction, which means a value less than π and that will transform the second half of X_{kf} in the negative frequency components ranging from minus 0.5 the Nyquist frequency to slight below zero. In order for this to be true, the time signal has to be low-pass filtered at half the sampling frequency.

The discrete Fourier transform, requires a high number of complex mathematical operations. For instance, a signal constituted by N discrete data points, requires N^2 mathematical operations. When the Discrete Fourier Transfrom (DFT) was firstly numerically implemented, back in the 60's, the computational resources were poorly developed and this represented a great computational challenge. The FFT requires a number of operations equals to $N \log_{10} N$, contrary to the N^2 operations of the DFT. The only requirement from this new approach was that the number of points N, should be a power of 2, such as 128, 256, 512 and so on. The FFT is therefore, a simply and efficient algorithm for calculating the DFT equations. It is based in the equation (4.33) in its simplest version, also known as radix 2 algorithm. The detailed method to calculate the FFT is explained in the appendix A.

4.1.5.2 Windowing

Before applying the Fourier transform, the signal must be processed with a technique called windowing in order to minimize signal leakage effects. This method is equivalent to multiply the signal sample by a window function of the same length. In a simple way, this prevents the waveform to get truncated at the start and end. Because the signal is captured at sampled fixed intervals, the results obtained can vary with the location of the sample with respect to the waveform period [11], which may result in data discontinuities.

The windows can be considered as a way to fill the discontinuities in the data, by forcing the sampled data to zero at the beginning and end of the sampling period, making the sampled period appear to be continuous. The leakage error occurs when the signal is not windowed and is discontinuous. The FT treats the discontinuities as varying frequencies which cause them to appear as sidebands in the spectrum. Windowing affects the ability to resolve closely spaced frequencies, but on the other hand it harms the amplitude accuracy, so an optimal compromise must be achieved [11].

There are currently several window functions which can be applied to the signal. The most commons ones are the rectangular, flat top, Hanning, Hamming, Kaiser Bessel, Blackman and Barlett. The most used are the first three. Each one of these windows will return better or rougher results according to the purpose of the analysis. For most applications, the best solution pass by processing the data in different ways, although a Hanning window returns an optimum compromise as it provides good amplitude resolution of the peaks between bins, and at the same time minimal increment of the peak. A comparison between the Hamming(a), Barlett(b), Hanning(c) and Blackman(d) is present in figure 4.10, adapted from Scheffer.

- **Hanning** - The Han window was suggested by Julius von Hann [32] and is given by equation (4.35).

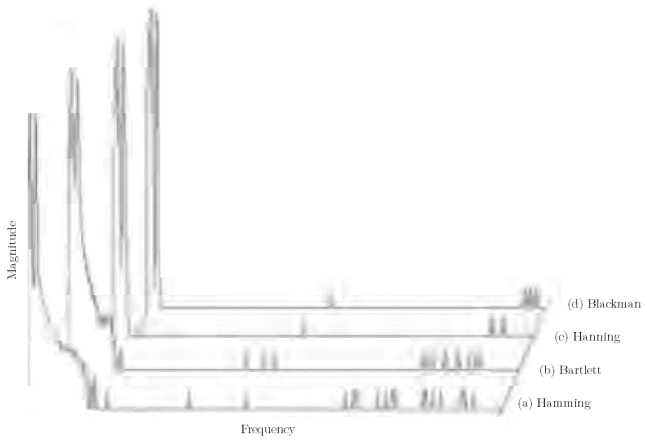


Figure 4.10: Comparison of different windows [11]

$$\mathcal{W}_{Hann}(n) = 0.5\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right] \tag{4.35}$$

- **Hamming** - another coefficients for the window were proposed by Richard Hamming. This window is optimized to minimize the maximum nearest side lobes [33].

$$\mathcal{W}_{Hamming}(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{4.36}$$

- **Blackman** - for the Blackman choice, other windows coefficients are adopted which place zeros at the third and fourth side lobes [34].

$$\mathcal{W}_{Blackman}(n) = a_0 - a_1\cos\left(\frac{2\pi n}{N-1}\right) + a_2\cos\left(\frac{41pin}{N-1}\right) \tag{4.37}$$

$$a_0 = \frac{1-\alpha}{2} \qquad a_1 = \frac{1}{2} \qquad a_2 = \frac{\alpha}{2} \tag{4.38}$$

- **Bartlett** - is also known as a triangular window. The Fourier transform of the result returns the squared values of the transform for half-width rectangular window [32].

$$\mathcal{W}_{Bartlett}(n) = 1 - \left|\frac{n - \frac{N-1}{2}}{\frac{N-1}{2}}\right| \tag{4.39}$$

- **Flat Top** - is defined as a partially negative valued window that as a limit in the frequency domain. This window has a low amplitude measurement error, however this is accomplished harming the frequency resolution. This problem can be minimized by a longer window length [35].

$$\begin{aligned}\mathcal{W}_{Flat}(n) = a_0 - a_1 \cos\left(\frac{2\pi n}{N-1}\right) + a_2 \cos\left(\frac{4\pi n}{N-1}\right) \\ - a_3 \cos\left(\frac{6\pi n}{N-1}\right) + a_4 \cos\left(\frac{8\pi n}{N-1}\right)\end{aligned} \tag{4.40}$$

Some properties of the Discrete Fourier transform are listed below [36], where $X(e^{j\omega})$ represents the discrete-time Fourier transform of $x(n)$.

1. **Linearity**
$$\mathcal{F}[\alpha\ x_1(n) + \beta\ x_2(n)] = \alpha\ \mathcal{F}[x_1(n)] + \beta\ \mathcal{F}[x_2(n)] \tag{4.41}$$
2. **Time Shifting**
$$\mathcal{F}[x(n-k)] = X(e^{j\omega})e^{-j\omega k} \tag{4.42}$$
3. **Frequency Shifting**
$$\mathcal{F}[x(n)e^{j\omega_0 n}] = X(e^{j(\omega-\omega_0)}) \tag{4.43}$$
4. **Conjugation**
$$\mathcal{F}[x^*(n)] = X^*(e^{-j\omega}) \tag{4.44}$$
5. **Folding**
$$\mathcal{F}[x(-n)] = X(e^{-j\omega}) \tag{4.45}$$

4.1.6 Time-Frequency Analysis

A time-frequency representation uses a 2-axis coordinate system to plot the frequency and time giving the magnitude as colour. There are several techniques performing this representation, which show potential to detect and diagnose bearing problems in complex rotating machines, where the SNR is very low and in the presence of several frequency components.

4.1.6.1 Short Time Fourier Transfrom

The Short Time Fourier Transfrom (STFT) is one of the most widely used time-frequency techniques [24]. The STFT is obtained by applying a windowing function to the original time signal and evaluating the conventional Fourier transform of the resulting finite length time signal [1]. In other words, this simply means to move a short time window along the record, an obtain the Fourier spectrum as a function of time shift. The STFT, is given by equation (4.46) where τ is the time shift and $\mathcal{W}(t)$ is the window moved along the signal.

$$S(f,\tau) = \int_{-\infty}^{\infty} x(t)\ \mathcal{W}(t-\tau)\ e^{-j2\pi ft}\ dt \tag{4.46}$$

The STFT is useful to track changes in frequency with time. However, due to the uncertainty principle, the frequency resolution is the reciprocal of the effective time window

length. The window can have finite length or theoretically infinite, as a Gaussian window. Regardless this fact, the signal will always have to be truncated. It is usual to display the amplitude squared on a time-frequency domain $|S(f,\tau)|^2$ which is known as spectrogram [1].

4.1.6.2 Wigner-Ville Distribution

The Wigner-Ville Distribution (WVD) is used on time-frequency analysis to eliminate interference between positive and negative frequency components. This distribution was initially proposed by Wigner and later modified by Ville [1]. The operation is part of a time-frequency distribution called Cohen class, which can be defined by the expression (4.47). The WVD have been used over the past 70 years in areas as radar and sonar, being particularly interesting in CM, for diagnosing short time non-stationary events such as impacts, amplitude and phase demodulation.

$$C_x(t,f,\phi) = \mathcal{F}[WVD(t,\tau)] \tag{4.47}$$

This distribution can be defined as a weighted autocorrelation function, defined by equation (4.48), where $\phi(u,\tau)$ is a kernel function allowing the smoothness of the WVD that is obtained for $\phi = 1$ [1]. The time signal must be known a priori for all time, which may not occur. To overcame this drawback, it is possible to apply the analysis to a windowed version of the analytic signal which is known as Windowed Wigner-Ville Distribution (WWVD) or Pseudo Wigner-Ville Distribution [24]. This improvement, allows the interferences in time and frequency to be suppressed and is computed over a time length T instead of all the domain, as can be seen on equation (4.49). This technique, gives better resolution than the STFT in both directions than the STFT.

$$WVD(t,\tau) = \int_{-\infty}^{\infty} x\left(u+\frac{\tau}{2}\right)\ x^*\left(u-\frac{\tau}{2}\right)\ \phi[(t-u),\tau]\ du \tag{4.48}$$

$$W_{WVD}(t,\tau) = \int_{-T}^{T} x\left(u+\frac{\tau}{2}\right)\ x^*\left(u-\frac{\tau}{2}\right)\ \mathcal{W}(\frac{\tau}{2})\ \mathcal{W}^*(-\frac{\tau}{2})\ \phi[(t-u),\tau]\ du \tag{4.49}$$

4.1.7 Hilbert Transfrom

In contrast to what happen with other integral transforms, where there is a change in the domains, such as the Fourier or Laplace transformations, the Hilbert Transfrom (HT) process a function g in another one on the same domain. This operation, assigns a complementary imaginary part to a given real part or vice versa, by shifting each component of the signal by a quarter of a period [37]. In other words, it can be said to be the relationship between the real and imaginary parts of the FT of a one-sided function [1]. Therefore, the HT provides a method for determining the instantaneous amplitude and frequency of a signal. The initial studies with this transform were conducted by David Hilbert in the beginning of the XX century but only in 1925 it was mathematically established by Godrey Hardy. Until 30 years ago, the HT was only a theoretical approach and it had not returned results that proved its utility. The only results have been received in the last 15 years, where the transformation was applied in research studies [37]. The Hilbert Transform $\mathcal{H}[g(t)]$ of a signal $g(t)$ is defined [38] by the equation (4.50).

$$\begin{aligned}\mathcal{H}[g(t)] &= g(t) \times \frac{1}{\pi t} \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{g(\tau)}{t-\tau}\, d\tau \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{g(t-\tau)}{\tau}\, d\tau\end{aligned} \tag{4.50}$$

The HT of $g(t)$ is the convolution of $g(t)$ with the signal $1/\pi t$. This is the response to $g(t)$ of a linear time-invariant filter called Hilbert transformer, having impulse response $1/\pi t$. The integral given by equation (4.50) is improper, because the integrand has a singularity and the limits of integration are infinite. The HT is properly defined by the Cauchy principal value when it exists. This value is obtained by considering a finite range of integration that is symmetric about the point of singularity, excluding at the same time the symmetric subinterval. This is achieve by taking the limit of the integral when the length of it tends to ∞ and the length of the excluded interval tends to zero [38]. This is easily understandable by equation (4.51).

$$\mathcal{H}[g(t)] = \frac{1}{\pi}\lim_{\epsilon\to 0^+}\left(\int_{t-1/\epsilon}^{t-\epsilon} \frac{g(\tau)}{t-\tau}\, d\tau \;+\; \int_{t+\epsilon}^{t+1/\epsilon} \frac{g(\tau)}{t-\tau}\, d\tau\right) \tag{4.51}$$

Some properties of the Hilbert transform are listed below [38].

1. **Linearity**
$$\mathcal{H}[a_1 g_1(t) + a_2 g_2(t)] = a_1\mathcal{H}[g_1(t)] + a_2\mathcal{H}[g_2(t)] \tag{4.52}$$
2. **Constant Signal** - For any constant c the transformation is 0.
$$g(t) = c \qquad \mathcal{H}[g(t)] = \mathcal{H}[c] = 0 \tag{4.53}$$
From the linearity property enumerated before it came that
$$\mathcal{H}[g(t) + c] = \mathcal{H}[g(t)] + \mathcal{H}[c] = \mathcal{H}[g(t)] \tag{4.54}$$
3. **Time Shifting** If $g(t)$ has Hilbert transform $\mathcal{H}[g(t)]$ then the transformation of $g(t-t_0)$ is given as follows
$$\mathcal{H}[g(t-t_0)] \tag{4.55}$$
4. **Convolution**
$$\mathcal{H}[g_1(t) \times g_2(t)] = \mathcal{H}[g_1(t)] \times g_2(t) = g_1(t) \times \mathcal{H}[g_2(t)] \tag{4.56}$$

According to Randall [1], the Hilbert transform can be achieved more simply by transforming the signal into the frequency domain, shifting the phase of positive frequency components by $-\pi/2$ and of the negative frequency components by $+\pi/2$ and returning to time domain after the shifting. One of the advantages of the HT is the fact that the function $g(x) = f(x) + j\mathcal{H}[u(x)]$ always represent an analytic function, also known as regular function in the superior half of the complex domain. This means that the function is infinitely differentiable. Despite other useful applications, nowadays the biggest advantage of the HT is the ability to obtain the envelope by demodulation of a signal.

4.1.8 Wavelet Transform

The concept of wavelets started to appear frequently in the early 1980's [39]. The wavelet transform is another specific type of time-frequency analysis that decompose the signal in a family of wavelets which have a fixed shape but can be shifted and dilated in time [1]. It is similar to STFT and the WVD as it provides a time-frequency map of the signal being analysed [24]. This new concept is the combination of different disciplines including mathematics, physics and engineering. The first approach was taken by Jean Morlet, who was a french engineer. Morlet discovered the idea of the wavelet transform as a mathematical tool for seismic wave analysis. The basis for his concept was that a signal consist of different discrete values in time and frequency, however, their high frequency components would have a shorter time duration than their low frequency components. The formula for the wavelet transform [1] is stated in equation (4.57).

$$\Psi(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \Psi^* \left(\frac{t-b}{a} \right) \, dt \tag{4.57}$$

Where $\Psi(t)$ is the mother wavelet, translated by b and dilated by a factor a. In other words, this means that the parameter a measures the degree of compression and b the time location of the wavelet. The basis for this approach was to provide better time resolution to the high frequency components and less to the low ones. This happens when $|a| < 1$, which means that the wavelet is the compressed version of the mother wavelet and corresponding to higher frequencies. In contrary, when $|a| > 1$ the wavelet functions have a larger time width than the mother wavelet, corresponding to lower frequencies. Therefore, it can be said that the wavelets have time-widths adapted to their frequencies, however the resolution of wavelets at different scales varies in the time and frequency domains, according to the principle of uncertainty of Heisenberg.

The wavelets can be considered as a set of impulse responses of filters, which because of the dilation factor maintain the bandwidth properties. Their main use is to detect local faults in gears and bearings [1]. The wavelets can be orthogonal or non-ortoghonal and continuous or discrete. Currently there are several wavelet families such as the Haar, Daubechies, Meyer, Gaussian, Mexican hat, Morlet, Shannon, Fejer-Korovkin among others. The wavelet can be complex with the advantage that the imaginary part of the wavelet is orthogonal to the real part, which allows the overall result to be independent on the position of the event being transformed. Orthogonal wavelets are the most efficient to use, if the analysis is to be performed after denoising or if important features are to be represented with a minimum number of parameters. For analytical purposes, non-orthogonal wavelets, such as Morlet, are often more convenient [1]. The wavelet transform of f is defined by equation (4.58) where $\Psi(a,b)$ act has the kernel polynomial in the Fourier transform. Similarly to the Fourier transformation, the continuous wavelet transformation $\mathscr{W}_\Psi$ is linear.

$$\mathscr{W}_\Psi[f](a,b) = (f, \Psi(a,b)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \; \Psi \left(\frac{t-\bar{}b}{a} \right) dt \tag{4.58}$$

The inverse wavelet transform can be defined, allowing f to be reconstructed [39], by the formula (4.59).

$$f(t) = C_{\Psi}^{-}1 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathscr{W}_{\Psi}[f](a,b)\Psi(a,b)(a^{-2}da)db \tag{4.59}$$

Where C_{Ψ} has to respect the admissibility condition given by equation (4.60), where $\mathcal{F}[\psi(\omega)]$ represents the Fourier transform of the mother wavelet $\psi(t)$.

$$C_{\Psi} = 2\pi \int_{-\infty}^{\infty} \frac{\mathcal{F}[\psi(\omega)]^2}{|\omega|} d\omega \ < \ \infty \tag{4.60}$$

The primary application of wavelets in machine diagnostics is in denoising signals in time and frequency domain simultaneously, as explored in the reference [40]. Wavelet denoising usually define two types of tresholding to remove noise, which are defined as any components with amplitude less than a certain threshold value. In the first technique, also known as hard thresholding, the retained components are left unchanged while in the soft thresholding, the noise estimate, defined by the threshold value, is subtracted from the retained components. As had been said, there are several wavelet families which perform better according to the needs of the application in study. The main criteria to chose the wavelet family is the similarity to the features in the signal which have to be extracted. For bearing faults, which generate impulsive signals, the impulse wavelet may return better results. Despite this fact, it is common to chose the wavelet family by trial and error or often based on the knowledge of the analyst.

Morlet Wavelets

Due to its leadership in this field, the Morlet Wavelets will now be former explored. This wavelet family is constituted by Gaussian windowed sinusois which are non-orthogonal. Those are used to analyse several vibration signals due to their similarity in appearance to the narrowband impulse responses. It is possible to tune this wavelets to evaluate different damping coefficients. The complex Morlet wavelet is defined in the time domain as a complex exponential wave multiplied by a Gaussian function, as seen in equation (4.61), retaining the same shape as the last. The original Morlet wavelets were windowed as cosines but it is convenient to use the complex version, with a one sided spectrum, which enables the imaginary part to be the Hilbert transform of the real part, as in a windowed sine wavelet [1].

$$\Psi(t) = \frac{\sigma}{\sqrt{\pi}} \exp(j2\pi f_0 t - \sigma^2 \ t^2) \tag{4.61}$$

Where f_0 represents the centre frequency of the window and σ its width. It is possible to obtain the mother wavelet $\Psi(f)$ by performing a FT using equation (4.62) of $\Psi(t)$, where $\psi^*(f)$ denotes the complex conjugate of $\Psi(f)$ that is equal to itself because the mother wavelet is real. The advantage of use the complex Morlet wavelets compared with the original cosine windowed ones is that the imaginary parts, have their maximum value when the cosines have zero crossings, which means that the squared amplitude of the wavelet coefficients is independent of the features local phasing in the time signals [1]

$$\Psi(f) = \Psi^*(f) = e^{-[\frac{\pi}{\sigma}(f-f_0)]^2} \tag{4.62}$$

4.1.9 Demodulation

The modulation phenomena occurs when a sinusoidal signal, also known as carrier signal has its amplitude or frequency varying with time. When the amplitude is variable on time it is called amplitude modulation and when is the frequency it is called frequency or phase modulation. Phase modulation is the deviation in the phase from the linearly increasing phase of the carrier wave while frequency modulation is the difference in velocity from the constant carrier frequency [1]. In other words, it can be said that frequency modulation is the derivative of phase modulation. There are several modulation phenomenons that occurred in mechanical applications such as shaft torsional vibration, meshing frequencies in a gearbox or race faults in rolling element bearings.

The rolling element bearings are the case of interest so lets consider a defective bearing with a single spalling on the outer race. Every time a ball passes through that spall it will generate an impact. Lets now assume that for each revolution, 5 balls pass through the spall, which means that there are 5 impacts generated for each shaft revolution. If the shaft is rotating at 1000 rpm, this means that there will be 5000 ipm has in impact per minute. After performing a FFT there will be a component in the 83 Hz which corresponds to 5000 ipm. However, this single peak in the spectrum may go undetected because it is located in the low frequency range of the spectrum which is generally smeared with noise frequencies [11]. To solve this problem, the envelope technique, which will be seen with more detail afterwards is used to diagnose faults in bearings. When the ball hit the spall, the bearing resonates due to the excitation induced from the impact. As had been said, the resonance frequency for bearings is located in the high frequency range [11]. When the bearing is rotating, there is the generation of a frequency of rolling elements hits on the outer race, previously defined as BPFO and the bearing is continuously being excited and resonating. This resonance reaches the maximum value when the ball hit the fault and starts to decrease until the next ball excite it again, which perform the modulation of the signal.

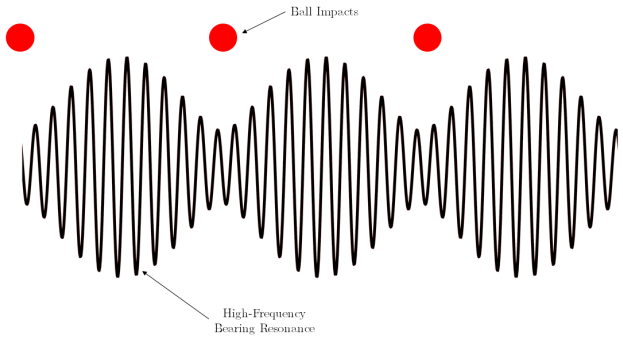


Figure 4.11: Modulation scheme for outer race fault [11]

It is obvious that the frequency of interest is the repetition of the impacts generated by the balls and not the resonance frequencies. The demodulation, clears the signal from the high frequencies present. The generally modulated signal can be represented by the expression (4.63), where $A_m(t)$ represents the amplitude modulation function and $\phi_m(t)$ represents the phase modulation function in radians. The corresponding frequency modulating function is given in Hz by the equation (4.64).

$$A_m(t)\cos(2\pi f_c t + \phi_m(t)) \tag{4.63}$$

$$\frac{1}{2\pi}\frac{d\phi_m(t)}{dt} \tag{4.64}$$

The equation (4.63) represents the real part of a rotating vector whose modulus is the amplitude modulating function, plus a DC offset. To ensure that $A_m(t)$ is always positive, which phase is the phase modulating function plus the linear carrier component [1]. Demodulation is mostly used in communication systems, where shifting of a frequency bandwidth by a high-frequency carrier, plays an important role in signal transmission. To see how the Hilbert transform is used lets focus on figure 4.12 which shows an amplitude modulated wave with a carrier frequency of 20 Hz and a modulated frequency of 5 Hz. On the right, there is the spectrum of the signal, where the sidebands around the carrier frequency are notorious

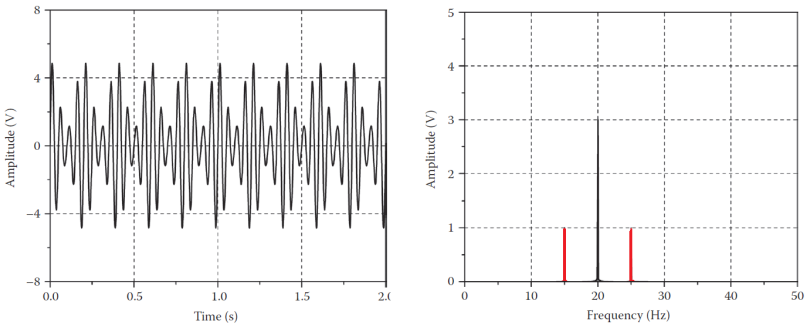


Figure 4.12: Time(a) and frequency(b) domain of a modulated wave [7]

Applying the HT, the envelope of the amplitude modulated wave is obtained and represented in figure 4.13. On the right, the 5 Hz is now visible in the spectrum.

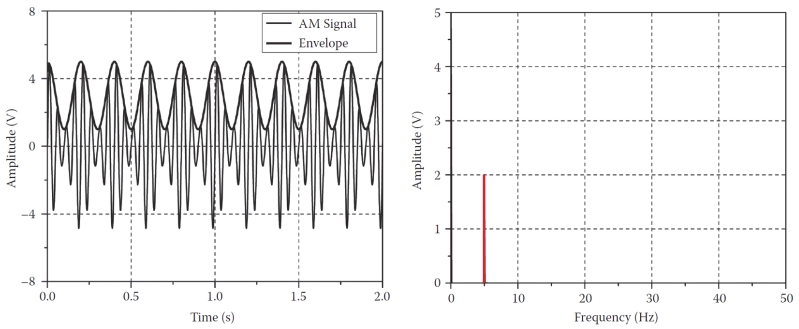


Figure 4.13: Envelope(a) and spectrum(b) after the Hilbert Transform [7]

The sidebands in figure 4.12 are spaced at the modulated frequency (5 Hz). In a coordinate system, rotating with the carrier, the upper sideband rotates in the positive direction at the modulating frequency, while the lower sideband rotates at the same frequency in the opposite direction. The phase relationships of the sidebands must respect the condition that their vector sum is always aligned with the carrier component enabling the change of its amplitude sinusoidally. If the phase is reversed, the modulation is in phase instead of amplitude.

4.2 Diagnostic Techniques

4.2.1 Minimum Entropy Deconvolution

Minimum Entropy Deconvolution (MED) was originally proposed by Wiggins for application on seismic data, as a mean to sharpen reflections from different subterranean layers, back in 1978 [41]. The aim of this method is to separate the components of a signal. This selects iteratively a Finite Impulse Response (FIR) filter to minimize the entropy of the filtered signal and has been used across various fields. The MED algorithm, purpose to extract the fault impulses while minimizing the noise. Therefore achieving a clear detection results even in the presence of high noise. The method, reduces the randomness of a signal my minimizing the entropy as cleared in the name [42].

This technique reduces the transmission path effects, from the source to the transducer, verified in the vibration signals measured externally on machines. In other words, it can be said that the MED is designed to reduce the spread of the Impulse Response Function (IRF) to obtain signals closer to the original impulses that originate them. This is quite common on signals originated by internal sharp impacts such as spalls in bearings, which have a strong impulsive nature [1]. The most common diagnostic techniques rely on being able to identify several response pulses providing from the impacts between the balls and the races. This will only be achieved if the IRF are shorter than the spacings between them, which may not happen for high speed equipments. This section will firstly present the MED algorithm and then the autoregressive model, used by Endo and Randall in 2007, in conjugation with the MED.

4.2.1.1 Algorithm

The increasing entropy in a signal translates in higher disorder. The more impulsive a signal is the more structured it will be, requiring all significant frequency components to have zero phase simultaneously at the time of each impulse. In other words, this means that minimizing the entropy corresponds to maximizing the structure of the signal, which consequently leads to higher impulsiveness and maximum kurtosis of the inverse filter output [41]. This corresponds to the original input in the system [1]. The principle of the MED is illustrated by the figure 4.14 where $d[n]$ represents the signal originated on the bearing and $u[n]$ other signals generated on the different components that may be present in the mechanical equipment. These two signals pass through the structural filter h whose output is mixed with noise $e[n]$ before being measured on the sensor as $x[n]$. The MED applies an inverse filter f to produce the output $y[n]$ which is expected to be as close as possible to the original signal generated on the bearing $d[n]$. Despite not being known, the input $x[n]$ for the algorithm is expected to be a very impulsive signal, with the method returning better results for the more impulsive signals.

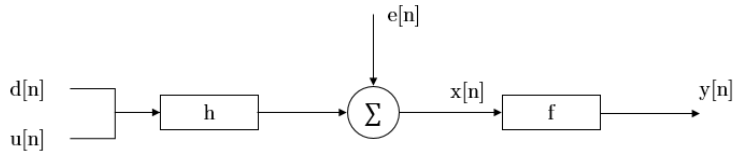


Figure 4.14: Deconvolution process for MED

Now it will be explained the process to perform the MED to a signal in a detailed manner. At the end of this section there is a compressed explanation which can be implemented into a *script*. Starting from a general linear time-invariant acceleration signal, proving from a machine, where x_n represents the sampled acceleration signal constituted by, u_n which is an unknown input sequence, d_n which is the repeating impulse fault signal input sequence and e_n, the noise. The scalars b_k and c_k represent the dependence of x_n on previous x_n, u_n and d_n respectively.

$$\begin{aligned} x_n = &(b_1 u_{n-1} + b_2 u_{n-2} + \cdots + b_L u_{n-L}) \\ &+ (c_1 d_{n-1} + c_2 d_{n-2} + \cdots + c_M d_{n-M}) + e_n \end{aligned} \tag{4.65}$$

Any stable transfer function can be approximated as a FIR filter. The transfer functions in every equipment are stable or partially stable otherwise the machine would have infinitely growing vibration and would in theory explode [41]. Due to this fact, the resulting time domain approximation form in terms of convolution is given by equation (4.67), where $\vec{h}$ are the FIR filter approximations to their respective transfer functions under stable assumption and N is the number of measured samples of the signal.

$$\vec{x} = \vec{h}_u \cdot \vec{u} + \vec{h}_d \cdot \vec{d} + \vec{h}_e \cdot \vec{e} \tag{4.66}$$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \qquad \vec{u} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \qquad \vec{d} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \tag{4.67}$$

The deconvolution problem aims to reconstruct the fault signal $\vec{d}$ by applying the FIR filter $\vec{f}$ with size L to the measured machine acceleration signal $\vec{x}$.

$$\vec{y} = \vec{f} \cdot \vec{x} \qquad \vec{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_L \end{bmatrix} \tag{4.68}$$

The goal is that the resulting filtered signal $\vec{y}$ approximates the fault signal $\vec{d}$ which is achieved by selecting the filter to minimize the noise effect and cancelling the effect of the system, also known as transmission path effect. The appropriate selection of the filter may be

difficult but the fault signal is expected to be an impulsive signal, with high kurtosis, while the remaining signals $\vec{u}$ and $\vec{e}$ have very low kurtosis. The filter is selected to reach the maximum kurtosis, as a result of the kurtosis difference between the signal components. The proposal made by Wiggins was to maximize a norm function called Varimax Norm, which in the case of the one dimensional MED is equivalent to maximizing the kurtosis, discussed in the Chapter 5. Assuming a zero mean, the maximum kurtosis is given by equation (4.69).

$$K_{Max}(\vec{f}) = \frac{\sum_{n=1}^{N} y_n^4}{[\sum_{n=1}^{N} y_n^2]^2} \tag{4.69}$$

This is also called the objective function method where the maximum kurtosis is obtained by varying the coefficients of the filter [1]. The coefficients for the filter are obtained by derivation the equation (4.69) with respect to themselves and solving it equal to zero as shown by equation (4.70).

$$\frac{dK_{Max}(\vec{f})}{d\vec{f}} = 0 \tag{4.70}$$

The filter is iteratively selected. The algorithm can be synthesised as follows

- Step 1: Assume initial filter as a centred impulse, i.e $\vec{f} = [\ 0\ 0\ \cdots\ 1\ \cdots\ 0\ 0\ ]^T$.
- Step 2: Calculate $\vec{y} = \vec{f}\vec{x}$.
- Step 3: Determine the new filter coefficients by solving for $\vec{f}$ in equation (4.70).
- Step 4: Repeat from step 2 for a specified number of iterations or until the change in kurtosis between iterations is below a specified value.
- Step 5: Obtain the final deconvolved signal.

The biggest disadvantage of the MED technique is that with a finite length signal, composed only by white noise, the method is able to successfully deconvolve a single impulse for a sufficiently large filter. This is commonly seen as a problem since in the ideal no-fault cause the kurtosis should be close to 3 and the resulting single impulse may be of higher kurtosis which may lead the analyst to improperly diagnose a fault state. In addition to that, the algorithm maximizes the kurtosis as a solution with the fewest number of impulses which in some cases may lead to deconvolved solutions with fewer than the desired impulses [41].

4.2.1.2 Autoregressive Model

The autoregressive models have been gaining importance in the field of rotating machines vibration fault detection. As declared by [41], this model has been effective in extracting gear faults with no previous knowledge of the fault. The model with no input is given by equation (4.71) where a_i are scalar model coefficients, N is the order of the model, e_n is the with noise and y_n is the signal being modelled.

$$y_n = -a_1 y_{n-1} - a_2 y_{n-2} - \cdots - a_N y_{n-N} + e_n \tag{4.71}$$

This means that the current sample is a linear combination of the N previous samples plus an additional white noise. To determine the scalar model coefficients there are several approaches. This algorithm used the Burg's lattice based method due to the robustness of the estimation. This approach selects the parameters by minimizing the least-squares of both forward and backward prediction errors. The implementation is performed as follows.

- Step 1: Select an autoregressive model of order N using an appropriate criteria such as the Akaike information one.
- Step 2: Fit the model to the no-fault data by calculating the scalar coefficients with the Burg's method.
- Step 3: Return to step 1 after the prediction on the faulty vibration data and calculate the prediction error.

The prediction error consists of white noise, disturbances and potentially important data about the fault. This is one of the several disadvantages of this method. This means that, in order to perform well, the fault signal must be significantly larger in amplitude than the noise e_n. In addition to that, it is necessary to have the knowledge of the system in the healthy condition [41]. The combination of the autoregressive model and the MED is sometimes mentioned as ARMED. Together, those techniques allows the autoregressive operation to perform a spectrum whitening, but due to the limitation of being an autocorrelation function it has no phase information, then the MED achieves the phase alignment to maximize the impulsiveness of the filtered signal. One example of the application of this technique is given in figure 4.15 where it is shown the effect of the steps of the method in the time signal.

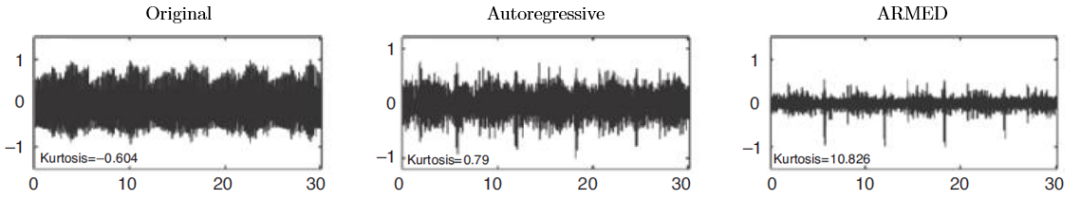


Figure 4.15: The effect of autoregressive plus MED method for gear fault detection [1]

4.2.2 Maximum Correlated Kurtosis Deconvolution

An improved deconvolution norm, known as Correlated Kurtosis (CK) takes advantage of the periodicity of the faults to achieve better deconvolved signals. This deconvolution technique, defined as Maximum Correlated Kurtosis Deconvolution (MCKD) propose to select a FIR filter to maximize the CK of the resulting signal which emphasizes high kurtosis while encouraging periodicity about a specific period. In this text, the *M-shift* CK will be the case study, however for the first shift CK the reader could use [41]. The correlated kurtosis of *M-shift* is given by equation (4.72).

$$CK_M(T) = \frac{\sum_{n=1}^{N}(\prod_{m=0}^{M} y_{n-mT})^2}{(\sum_{n=1}^{N} y_n^2)^{M+1}} \tag{4.72}$$

$$y_n = \sum_{k=1}^{L} f_k x_{n-k+1} \qquad x_n, y_n = 0 \; for \; n \neq 1, 2, ...N \tag{4.73}$$

Where N is the number of samples in the input signal, L is the length of the FIR filter and T the period of interest. When $T = 0$ and $M = 1$ the CK is the kurtosis norm used by MED given by equation (4.69). For that case, the proposed CK reaches a maximum for a periodic impulse about the specified period, as opposed to the kurtosis which tends to a maximum with a single impulse as said before [41]. Using higher shifts in the correlated kurtosis, larger sequences of impulses will be emphasized in a row. The method starts with the maximization equation expressed in (4.74).

$$\frac{d}{d\vec{f}} CK_M(t) = 0 \tag{4.74}$$

The previous equation should be solved in order to the filter coefficients f_k, being the numerator and denominator given by equation (4.75) and (4.76) respectively.

$$\frac{d}{df_k} CK_{Numerator} = 2 \sum_{n=1}^{N} \left[\left(\prod_{m=0}^{M} y_{n-mT} \right)^2 \left(\sum_{m=0}^{M} \frac{x_{n-mT-k+1}}{y_{n-mT}} \right) \right] \tag{4.75}$$

$$\frac{d}{df_k} CK_{Denominator} = 2(M+1) \|\vec{y}\|^{2M} \sum_{n=1}^{N} y_n x_{n-k+1} \tag{4.76}$$

Combining the two previous equations and equalling to zero.

$$2 \; \|\vec{y}\|^{-2M-2} \sum_{n=1}^{N} \left[\left(\prod_{m=0}^{M} y_{n-mT} \right)^2 \left(\sum_{m=0}^{M} \frac{x_{n-mT-k+1}}{y_{n-mT}} \right) \right] \tag{4.77}$$

$$-2(M+1) \|\vec{y}\|^{-2M-4} \sum_{n=1}^{N} \left(\prod_{m=0}^{M} y_{n-mT} \right)^2 \sum_{n=1}^{N} y_n x_{n-k+1} = 0 \tag{4.78}$$

Converting to a matrix form with $k = 1, 2, ..., L$ and rearranging the results in the iterative solution.

$$\vec{f} = \frac{\|\vec{y}\|^2}{2|\vec{\beta}|^2} (X_0 X_0^T)^{-1} \sum_{m=0}^{M} X_m T \vec{\alpha_m} \tag{4.79}$$

Where X are the z-transform of x_n and $\vec{\alpha_m}$ and $\vec{\beta}$ are given by the equations below.

$$\vec{\alpha_m} = \begin{bmatrix} y_{1-mT}^{-1}(y_1^2 y_{1-T}^2 \cdots y_{1-mT}^2) \\ y_{2-mT}^{-1}(y_2^2 y_{2-T}^2 \cdots y_{2-mT}^2) \\ \vdots \\ y_{N-mT}^{-1}(y_N^2 y_{N-T}^2 \cdots y_{N-mT}^2) \end{bmatrix} \qquad \vec{\beta} = \begin{bmatrix} y_1 y_{1-T} \cdots y_{1-MT} \\ y_2 y_{2-T} \cdots y_{2-MT} \\ \vdots \\ y_N y_{N-T} \cdots y_{N-MT} \end{bmatrix} \tag{4.80}$$

For large M, around eight or more, the iterative method can result in loss of numerical precision because of exceeding the range of the floating point exponent. This is due to the fact that higher-shifts MCKD require better estimates of the fault period T and in practical reality, the period is a fractional number. An additional resampling stage should be introduced as a preprocessing step [41]. The user must take into consideration that the higher order shifts increase the complexity of the calculation and consequently the computational time. A condensed algorithm is expressed step-by-step below.

- Step 1: Select the period of interest T in the range of 20-300, in case it is not possible, downsample the dataset.
- Step 2: Calculate the z-transform from the input signal $\vec{x}$.
- Step 3: Select the filter size L and assume an initial centred impulse filter.
- Step 4: Calculate the filtered output $\vec{y}$.
- Step 5: Calculate the matrices $\vec{\alpha_m}$ and $\vec{\beta}$.
- Step 6: Calculate the new filter coefficients. If the number of iterations were reached or the increase in the CK was obtained terminate, otherwise return to Step 4.

4.2.3 Spectral Kurtosis

The Spectral Kurtosis (SK) was first introduced by Dwyer, back in the 1980's as a statistical tool which could indicate not only non-Gaussian components in a signal, but also their locations in the frequency domain [43]. In other words, this parameter provide a way to determine the frequency bands which contain the maximum impulsiveness. Initially it was used as a complement to power spectral density, to detect impulsive events in sonar signals which were strongly affected by noise. Originally it was based on the STFT and gave a measure of the impulsiveness of a signal as a function of frequency [1]. More precisely, it was defined as the normalised fourth-order moment of the real part of the STFT.

In the Chapter 5 there is a detailed description of the parameter Kurtosis as a global indicator. The difference between it and the SK is that the last one is applied locally in different frequency bands. In contrast to classical kurtosis analysis, the SK provides a robust way of detecting incipient faults even in the presence of strong masking noise and in addition to that, it offers an easy way of designing optimal filters for filtering out the mechanical signature of faults. This two advantages are of extreme importance for monitoring and diagnosis purposes respectively [44]. The definition proposed by Capdevielle was the SK as the normalised fourth-order cumulant of the Fourier transform, in other words as a slice of the tricoherence

spectrum. Although it perform well in the stationary case signals, the same was not true for non-stationary signals. The formalisation of the SK will be made by means of Wold-Cramér decomposition of conditionally non-stationary processes. This makes it possible to establish under which conditions the non-stationary process generates a non Gaussian distribution and can be described by time averages, or stationarised, statistics [43].

Assuming a fault signal $X(t)$ which has a transient nature, such as the incipient faults in rotating machines, one can claim the expression (4.81) where $h(t)$ is the impulse response resulting from a single impact and X_k,τ_k are sequences of random variables which account for possibly random amplitudes and random occurrences of the impacts respectively [44]. The reader should keep in mind that for any index k, it is assumed that the difference between τ_{k+1} and τ_k is a non-negative random variable whose mean specifies the average rate of repetition of the impacts.

$$X(t) = \sum_k h(t - \tau_k)\ X_k \tag{4.81}$$

As had been said the estimation of the SK is based on the STFT. For this estimator to be valid the analysed signal has to respect two conditions. The first one is that the non-stationarities in the signal have slow temporal evolutions as compared to the window length of the transformations and the second, the correlation length of the signal has to be shorter than the analysis window of the STFT. The kurtosis of each frequency f can be calculated by taking the fourth power of $H(t, f)$, as expressed in equation (4.82), which is the amplitude envelope function obtained by the STFT, at each time and averaging its value along the record normalizing it afterwards, by the square of the mean square value. The subtracted value 2 is to shift the result to zero for a Gaussian signal [1].

$$K(f) = \frac{\langle H^4(t, f)\rangle}{\langle H^2(t, f)\rangle^2} - 2 \tag{4.82}$$

The results obtained depend on the parameter chosen for the STFT, although the denominator of the equation is independent of the window length, the numerator is affected by it. To obtain the maximum value of kurtosis, the window must be shorter than the spacing between the pulses but longer than the individual pulses. There are some notes that should be taken into account when dealing with the concept of the SK. Its value decreases with the rate of repetition of the impulses, which means, with the rotation speed of the machine and when the window length of the STFT increase. On the other hand, the SK value increases with the sampling rate of the signal and with the intensity of the fluctuations in the impulse amplitude [44]. It is worth saying that a very short window length will produce SK with poor spectral resolution which may lead to some loss of information.

4.2.3.1 Detection Filters

The use of the SK as a filter is proposed by Randall [1] due to the fact that this parameter has a large value in frequency bands where the impulsive bearing fault signal is dominant and zero in frequencies dominated by stationary components. Typical detection filters used in vibration analysis are the Wiener filter, the matched filter and the dB-spectrum difference. The common aspect between all of them is that they can be blindly identified from the SK.

Wiener Filter is the optimal filter which recovers a signal $X(t)$ buried in additive noise $N(t)$. It is among all possible linear filter, the one that minimises the mean square error between $X(t)$ and $w(t)\cdot Y(t)$ where $w(t)$ is the filter and $Y(t)$ the signal measured. The solution for the Wiener filter is given by equation (4.83), where $\rho(f)$ is a signal to noise ratio value given by the power spectral density [44].

$$\mathcal{W}(f) = \frac{1}{1+\rho(f)} \tag{4.83}$$

In practice the exact solution for this equation is not achievable unless the value of $\rho(f)$ is known previously. This is only possible when some historical data are available, which is not the general case. This is where the SK is useful, giving access to the estimation of the Wiener filter without previous knowledge of the signal to noise ratio. This is done assuming that the SK is proportional to the square of the Wiener filter, equation (4.84) and the faulty signal can be recovered by simply filtering $Y(t)$ with the square root of the spectral kurtosis. This means that frequency components where the SK is high, will be enhanced in the signal, while the ones near to zero will be attenuated [44]. The scaling factor k does not compose a problem for recovering a scaled version of the fault signal but may be a barrier to determine the additive noise.

$$K(f) \approx k \times [\mathcal{W}(f)]^2 \tag{4.84}$$

Matched Filter in some situations it is preferable to obtain a filtered version of the signal which maximize the signal to noise ratio instead of recovering the exact shape of the transient signal. In this cases, the matched filter is solution, with the filter $\mathcal{M}(t)$ maximising the ratio of the power of $X(t)\cdot\mathcal{M}(t)$ to the power of $N(t)\cdot\mathcal{M}(t)$. The matched filter is given by the eigenvector associated with the largest eigenvalue of an autocorrelation matrix which has for elements, the inverse Fourier transform of $\mathcal{M}(f) = 1/\rho(f)$. Due to the relation with the Wiener filter, matched filter can be identified from the SK as expressed in equation (4.85).

$$\mathcal{M}(f) \approx \frac{\sqrt{K(f)/k}}{1-\sqrt{K(f)/k}} \tag{4.85}$$

The parameter k can be determined to maximise the kurtosis at the output. When the frequency corresponds to the maximum SK, the matched filter tends to an ideal band-pass filter tuned on the frequency where the signal to noise ratio is the highest.

Difference dB-Spectra is an empirical technique used in vibration monitoring that consists in comparing the dB-spectra before and after the fault has appeared. The difference between them indicates the frequency bands where the relative change of energy induced by the fault is the greatest and consequently where to band-pass the signal [44]. It is advisable to use this technique before the envelope analysis. The dB-spectrum difference is given by equation (4.86).

$$\Delta dB(f) = 10\log_{10}\left[1+\frac{1}{\rho(f)}\right] \tag{4.86}$$

The dB-spectrum difference takes values near zero where the signal to noise ratio is strong and high positive values where it is weak. It can be obtained from the SK by equation (4.87), which proves that the difference in dB-spectra is a monotonic function of the SK.

$$\Delta dB(f) \approx -10 \log_{10}[1 - \sqrt{K(f)/k}] \tag{4.87}$$

4.2.4 Kurtogram

The idea of the kurtogram is to create a robust detection filter $g(t)$ with an imposed band-pass structure with only two parameter to be identified. The objective of this technique is to find the central frequency f_c which tunes the filter where the SK is maximum, or in other words, where the signal to noise ratio is the highest, and the bandwidth B_f of the filter which achieves the best compromise between too wide filter, which would alter the signal to noise ratio, and too narrow filter with a very long impulse response that would alter the impulse like nature of the filtered signal [44]. Figure 4.16 shows the example of a kurtogram where N_W is the window length defining spectral kurtosis [1].

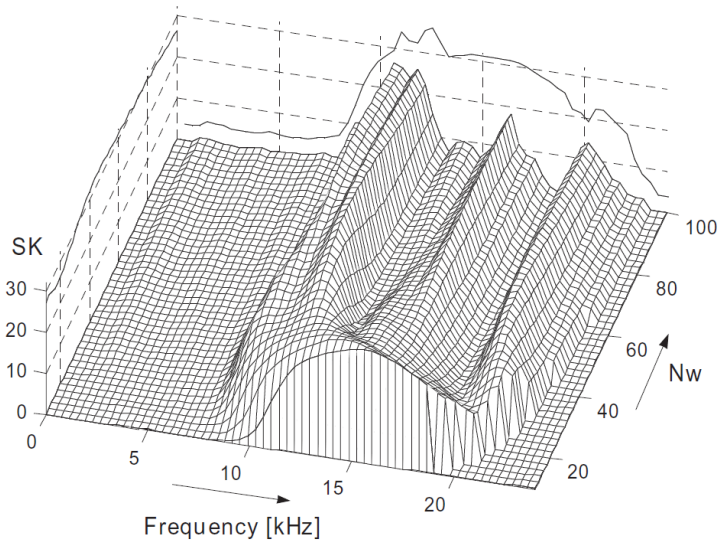


Figure 4.16: Example of a kurtogram for a outer race fault [1]

The basis beyond this approach is that a bandpass filter has good chance of selecting the frequency band where the signal to noise ratio is maximum. This could achieve detection of the one case with the strongest impulsiveness in cases with multi sources. It is indispensable to apply a band-pass filter before carrying out envelope demodulation. In addition to this, forcing a pre-defined structure on the filter makes it independent of the assumed validity of the model [44]. The computation of the full kurtogram covering all combinations of centre frequency and bandwidth is very costly and so a number of more efficient alternatives have been proposed. The most detailed one was the fast kurtogram based on a series of digital filters rather than the STFT. In the most basic version, the frequency range is progressively split into bands that

are one-half the width of the previous stage or scale, also known as binary tree. The version recommended however is the 1/3 binary tree where the split includes division into 1/3, so that the overall division is in the sequence 1/2, 1/3, 1/4, 1/6, 1/8, 1/12 and so on [1]. The resulting combinations of centre frequency and bandwidth are explained in figure 4.17.

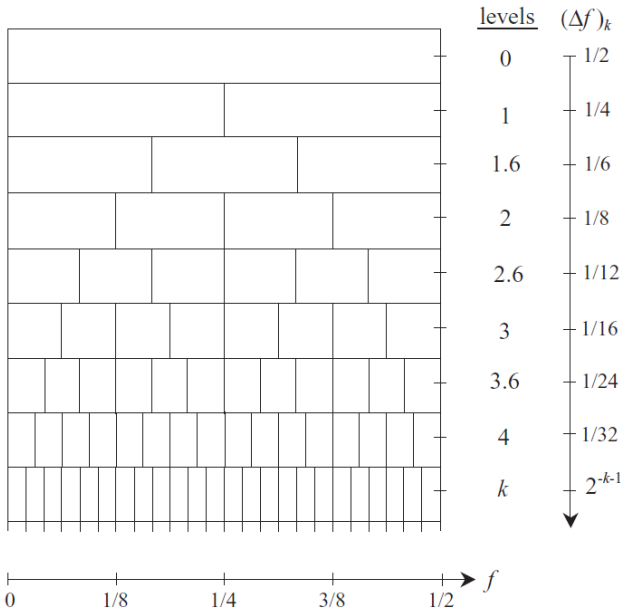


Figure 4.17: Representation of the 1/3 binary tree kurtogram [1]

Since the kurtogram is used to detect series of impulse responses such as the ones coming from bearing faults, and due to their approximate constant damping ratio, which manifest itself in the frequency domain as a constant percentage bandwidth structure, it can be said that combinations given by 1/3 binary tree are unlikely to be used and that a $1/n$-th octave wavelet analysis is more adequate for seeking the filter bandwidth and central frequency [1]. The discrete wavelet transform occupies fewer combinations than the fast kurtogram being limited to constant percentage bandwidth. Robert Randall proposes a wavelet kurtogram based on non-orthogonal complex Morlet wavelets in a publication made in 2005 with Nader Sawalhi.

4.3 Envelope Analysis

Envelope analysis is one of the most used techniques for detection and diagnosis of rolling element bearing incipient failure. It was developed in the early 1970's and was originally called the high frequency resonance technique. It has been known by other names including amplitude demodulation, demodulated resonance analysis, narrow band envelope analysis or just envelope analysis as we will adopt [24]. The fundamental basis to envelope analysis is the concept that each time a localised defect in a rolling element bearing makes contact, while under load, with another surface in the bearing, an impulse of vibration is generated. This impulse will have an extremely short duration compared to the interval between impulses and due to this, its energy will be distributed across a very wide frequency range. As a consequence to that, various resonances of the bearing and the surrounding structure will be excited by those impacts. These

excitations will normally be repetitive because the contacts between the defect and the mating surfaces will occur in a repetitive manner [24]. The frequency of the impulses are known as the fault characteristic frequencies and were studied in the Chapter 3 and it is usual to consider the resonance as being amplitude modulated at the characteristic defect frequency which makes it possible to detect the presence of a defect and its location.

The spectrum of the raw signal often contains little diagnostic information about bearing faults, with envelope analysis, the signal is bandpass filtered in a high frequency band in which the fault impulses are amplified by structural resonances. After that it is amplitude demodulated to form the envelope signal which spectrum contains the desired diagnostic information [10]. One of the biggest difficulties with envelope analysis is how to determine the best frequency band to envelope [24]. In this text, this band is chosen using the kurtogram, explained before. Randall and Antoni [10] state that considerable improvement can be made by taking advantage of digital processing techniques instead of following the analogue method in a digital form. In fact, according to them, a number of benefits arise from performing the amplitude demodulation using the Hilbert transform returning a complex time signal which imaginary part is the Hilbert transform of the real part, this means that the extraction of the part of the spectrum to be demodulated is performed by an ideal filter which can separate it from adjacent components that might be stronger than that, such as gearmesh frequencies in cases where gears are present. Figure 4.18 illustrates the procedure to perform envelope analysis.

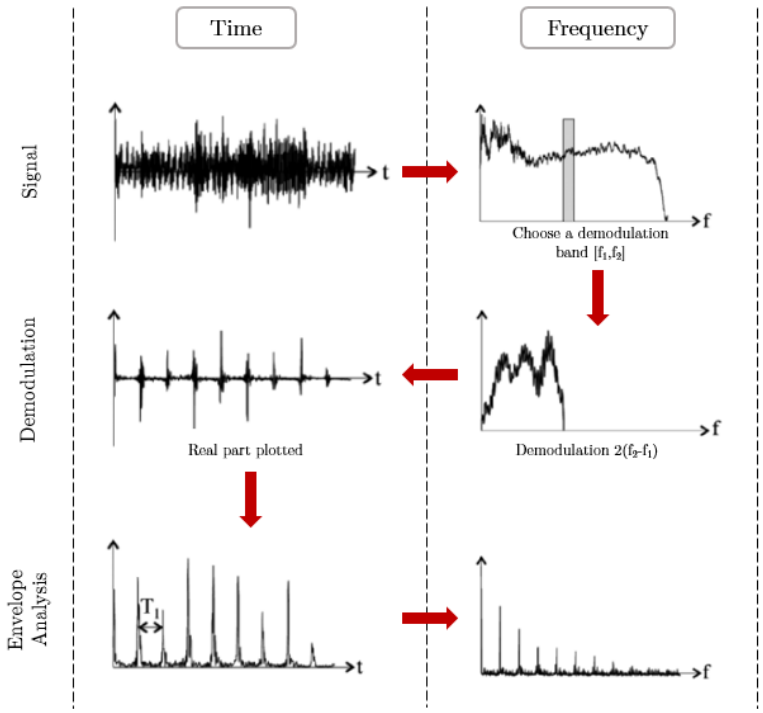


Figure 4.18: Procedure for envelope analysis [10]

According to Randall [1] it is preferable to analyse the squared envelope signal rather than the envelope, due to the fact that mathematically, the envelope of a signal is the square root of the squared envelope and therefore a rectified signal is the square root of the squared signal. The square root operation introduces external components that were not in the original squared signal and that will cause masking of the desired information. Since the beginning of the envelope analysis there has been a debate on how to choose the most suitable band for

demodulation. However, despite some authors propose the hammer test to find bearing housing resonances, it is commonly accepted that this problem can largely been solved by the use of SK and the kurtogram to find the most impulsive band, after properly remove the masking discrete frequency components [1].

4.3.1 Adopted Procedure

The method adopted to perform envelope analysis in this text was a combination of the techniques described in this chapter, minimizing the user intervention as much as possible. As detailed in figure 4.19 the process start with the discrete components removal. This is achieved using the Cepstrum editing method and allows the separation of the discrete frequencies and the random components, in other words it removes the gearmesh frequencies when gears are present and the shaft components in the single shaft case. The shaft speed is estimated by the tachometer signal and the parameter for the cepstrum should allow that 10-20 periods of the minimum frequencies are removed [1].

After this, we applied the MED to remove the smearing effect of the signal transfer path. Despite Randall [1] consider that this method only need to be applied for high speed bearings, where the impulse response of the bandpass filtered resonance has a length similar to the spacing of the bearing fault pulses. However, the best way to tune the parameter may be by trial and error based on the increase giving by the application of different parameters. After apply the MED the fast kurtogram is used to select the optimum band for demodulation. It should be taken into consideration that the kurtogram is sensitive to large random pulses which may be present in some parts of a signal. When the final envelope spectrum does not reveal periodic components, despite a high spectral kurtosis value, it should be checked if external impulses coming from other sources are dominant in certain frequency bands [1].

After the demodulation, the spectrum of the signal envelope is calculated and plotted It is obvious that the modulating effects are the most important part of the diagnosis. Inner race fault would be modulated at shaft speed and rolling element faults at cage speed. In situations with unidirectional load, the outer race fault would not be modulated. Unbalance and misalignment were not further investigated in this study however for those cases the modulation at the shaft speed can be visible and modulation at the cage speed can also result from variations between the rolling elements [1]. In the cases studied, the geometric properties of the bearings were known, so the fault characteristic frequencies were plotted against the envelope spectrum to assess if the fault was or not detected.

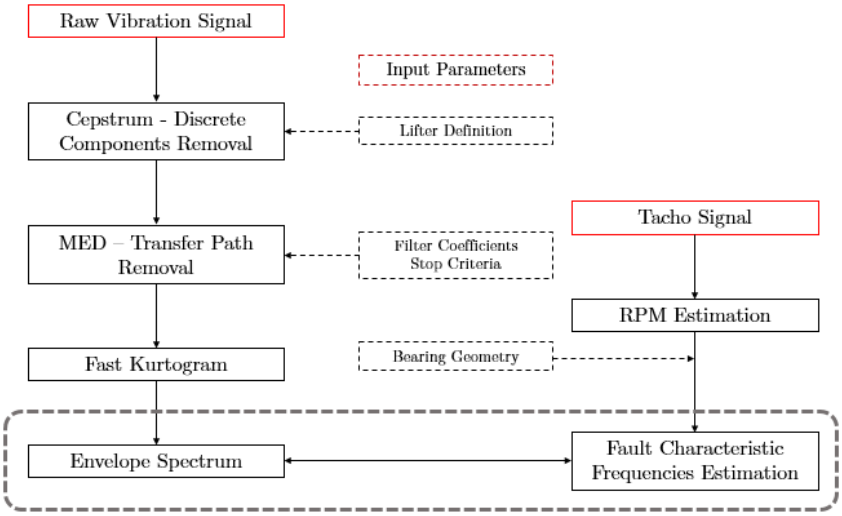


Figure 4.19: Procedure adopted for bearing diagnostics

Chapter 5

Features Extraction

The analysis of time domain parameters was already discussed in the Chapter 4. In this section, we will explain with more detail some of the features which are currently being used for fault detection and parameter trending. The success from computation and trending of time domain statistical parameters will depend on the mode of failure which is occurring and the class of vibration environment [24]. On this case is the peak vibration level, or in other words the maximum level, may be affected if a certain mode of failure influence a narrow frequency band. If the overall peak amplitude is not affected the fault can go undetected until the amplitude in the narrow frequency band becomes the largest component.

5.1 Statistical Parameters

Different approaches may be taken for time domain statistics. One of them consist in calculate the features for the whole frequency range of the signal and the second to divide the signal into discrete frequency bands and perform single analysis in each band.

The statistical parameters of a signal can be obtained from the probability density function by taking various moments. These are the moments of the curve and are analogous to mechanical moments about the centroid of a plane [4]. According to Davies the odd moments are related with information about the position of the peak density, relative to the median value, while the even moments indicate the spread in distribution. It is common to normalize the moments, that are higher than the second order, by removing the mean and dividing by the standard deviation raised to the order of the moment. The higher moments are sensitive to impulsiveness in the signal [4] which make moments such as the Skewness and Kurtosis, respectively third and fourth moments, so relevant to the diagnosis analysis.

5.1.1 Mean Value

The first moment of the probability density function corresponds to the mean value of the signal, because the area under the curve is unity and it defines the centre of gravity. As one can conclude, for a symmetrical function, the mean value will be at the line of symmetry as for instance the Gaussian function [1]. The mean, or expected value is given by equation (5.1).

$$\mu = \int_{-\infty}^{\infty} x \cdot p(x)\ dx \tag{5.1}$$

5.1.2 Variance

In the same way, the second moment of the probability density function is defined as variance. The variance allows the user to assess the spread of the signal from its mean. As it is known, the variance is the square value of the standard deviation. It is given by the equation (5.2).

$$\sigma^2 = \int_{-\infty}^{\infty} [x - \mu]^2 \; p(x) dx \tag{5.2}$$

5.1.3 Skewness

The third centred moment, returns a parameter called skewness which is a measure of the probability distribution asymmetry of a signal about its mean. The value of this feature may be positive or negative as the curve is right or left inclined respectively , as can be seen in figure 5.1. In other words, skewness is actually an indicator of the symmetry of the signal, from the CM point of view, the value of the skewness for good condition machinery should be around zero.

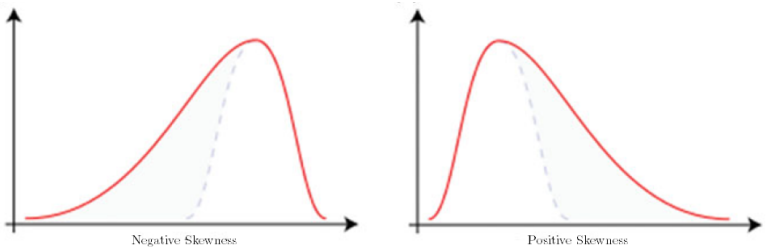


Figure 5.1: Example of negative and positive skewness [17]

In terms of vibrations, the higher asymmetry means higher irregularities in the time signal or the presence of defects in the machine. Because it is the third moment of the probability density function, the order of the power 3 enhance the high values and suppresses the low values in the signal, as expressed in equation (5.3). This fact make this parameter more useful for low-speed machinery although it is not widely used in industrial practice due to the need of high knowledge to understand the meaning of its values [17].

$$S = \frac{\int_{-\infty}^{\infty} [x - \mu]^3 \; p(x) dx}{\sigma^3} \tag{5.3}$$

5.1.4 Kurtosis

The kurtosis, which represents the fourth moment of the probability distribution function, is used to measure the impulsiveness or spikiness of the signal. Similarly to the skewness, this parameter is a descriptor of the shape of a probability distribution function [17]. Its value is very high for impulsive signals because of the weighting given to local spikes by the fourth power as seen in equation (5.4).

$$K = \frac{\int_{-\infty}^{\infty} [x - \mu]^4 \; p(x) dx}{\sigma^4} \tag{5.4}$$

For a Gaussian distribution, the value of the kurtosis is 3, so it is common to observe in some applications the use of a cumulant, where the value 3 is subtracted to the estimated value. A set of data with high kurtosis will produce a distribution curve with higher peak value at the mean and low values at the ends, which in other words means that there are more data at the extreme values from the mean [17].

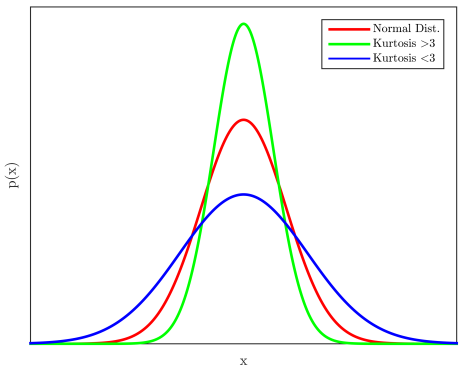


Figure 5.2: Different kurtosis values for probability density function

An high order cumulant is defined in a way which eliminates the components of lower order, which is not the case for general moments. For Gaussian signals, cumulants of order higher than second are equal to zero [1]. Figure 5.2 describes what is stated in the previous paragraph with higher peak values at the mean returning higher kurtosis values.

5.1.5 Peak-to-Peak

Peak-to-Peak (PP) is a feature used to measure the changes in the signal. Defining the peak as the maximum value of the signal and valley as the minimum value in the negative part of the y-axis.

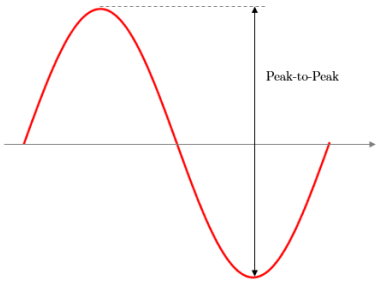


Figure 5.3: Peaks identification in a signal

Figure 5.3 intends to graphically represent this points for a sinusoidal signal. The peak-to-peak is calculated as the difference between them and is given by the equation (5.5). In a simple way it can be said that the peak-to-peak is the range between the maximum and minimum value in the signal.

$$PP = x_{Max} - x_{Min} \tag{5.5}$$

5.1.6 Root Mean Square

The RMS is a statistical measure of the magnitude of a varying quantity, which in terms of vibration, represents the overall vibration energy. The RMS is defined as the square root of the arithmetic mean of the squares of vector. For a vector x with length N its value is given by equation (5.6).

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} {x_n}^2} \tag{5.6}$$

Despite being a good initial indication of the vibration severity, for low-speed machines, this parameter should be used carefully essentially because of the low vibration energy of the impacts in this regime. In other words, relatively low RMS values may still implicate considerable deviations in the machine components movement [17]. To properly use this feature, the analysis should be focus on the evolution of the machinery condition with time, which means a historical trending must be recorded since the healthy condition of the equipment. Due to this constrain, isolated diagnosis are not recommended.

5.1.7 Crest Factor

The Crest Factor (CF) is an evaluation of the waveform described by the ratio of the peak value to an average value. In other words, it is an indication of the magnitude of the peaks in the waveform [17]. This feature provides an idea of how much impact is occurring in a time wave and at the same time is easy to calculate. It is given by the equation (5.7).

$$CF = \frac{|Peak|}{RMS} \tag{5.7}$$

When visualizing the spectrum as the FFT it is difficult to differentiate the impacts, which indicate bearing wear and faults, and random noise [17]. It is always advisable that the features are trended with time to allow the changes in the magnitude to be detected.

5.1.8 Entropy and Wavelet Entropy

The entropy is a quantitative measure of the system disorder. In other words, it can be stated that the entropy increases with the randomness of the signal . The concept of statistical

entropy was introduced by Claude Shannon as a concept of measuring the average missing information on a random source [45]. The Shannon entropy is given by equation (5.8) where x represents a N length signal.

$$Entropy = -\sum_{n=1}^{N} p(x_n) \cdot \log_2 p(x_n) \tag{5.8}$$

This feature, quantifies the unevenness of the probability density function and it takes the value of zero for a constant random variable. The parameters under study were the normalized Shannon Entropy (SE) and the Log Entropy (LE) which are given by equations (5.9) and (5.10).

$$SE = -\sum_{n=1}^{N} p(x_n)^2 \cdot \log p(x_n)^2 \tag{5.9}$$

$$LE = -\sum_{n=1}^{N} \log p(x_n)^2 \tag{5.10}$$

5.1.9 Impulse, Margin and Shape Factor

In 1992, Chun Qing Li developed a new approach using other non-dimensional parameters. The parameters studied were the Impulse Factor (IF), Margin Factor (MF) or clearance factor and the Shape Factor (SF). According to Howard [24] these three parameters were found to be useful under simulation conditions using a Gaussian probability density function as a model of fatigue spalling on bearings. The equations for impulse, margin and shape factor respectively are the following ones.

$$IF = \frac{x_{Max}}{\frac{1}{N}\sum_{n=1}^{N} |x(n)|} \tag{5.11}$$

$$MF = \frac{x_{Max}}{\frac{1}{N}\left(\sum_{n=1}^{N} \sqrt{|x(n)|}\right)^2} \tag{5.12}$$

$$SF = \frac{RMS}{\frac{1}{N}\sum_{n=1}^{N} |x(n)|} \tag{5.13}$$

The margin and impulse factors were the most useful in the studies conducted by Li, highlighting the margin factor as the most sensitive and robust on the detection of incipient spalling [24].

5.1.10 Percentile 75

The percentile is a statistical parameter used to indicate the value below which a certain percentage of observations remains when the signal is ordered ascending. Figure 5.4 shows the concept for the probability density distribution function.

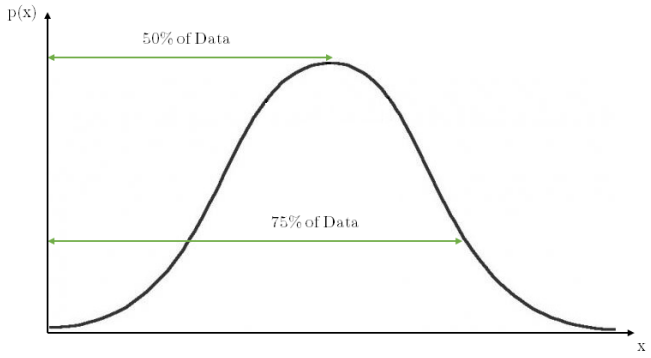


Figure 5.4: Graphical explanation of the percentile parameter

$$P_{75} = p\ (x < x_{0.75}) \tag{5.14}$$

5.1.11 Harmonics

Despite being conducted on the frequency domain. It was thought interesting to study the evolution of the harmonics amplitude with the operational conditions. The first 3 harmonics were studied and in order to normalize their value, the 0 Hz frequency magnitude was used. A coefficient was implemented due to the variations in the speed and load which may cause the harmonics to shifts from the theoretical values.

$$H_n = \frac{Max[S\ (n(1-r)f_T < f < n(1+r)f_T)]}{Max[\ S\ (0 < f < r)]} \qquad n = 1, 2, 3 \tag{5.15}$$

Where f_T represents the theoretic frequency, r the interval for searching, $S(f)$ the magnitude in the spectrum and n the n-th harmonic in study.

5.2 Features Analysis

As had been explained before, vibration signals generated by bearing faults are impulsive, at least at the source. However, the vibration signal measured at the bearing housing is the result of a combination of different types of signals determined by gears, transmission path and measurement noise. Shaft and gear related signals are deterministic and dominant and consequently they represent one of the most important masking components. The other

aspect to be taken into account is the transmission path from the source to the sensor which can decrease the impulsiveness of the signal.

5.2.1 Single Features

In the Chapter 6 the statistical features presented before, will be used to extract information about the presence of defects taking into account the speed and load on the shaft, in order to identify trends in those parameters and stablish limit levels. Another goal is to evaluate the benefits of different signal pre-processing techniques, such as the cepstrum editing and the MED. The procedure used to extracted the features is diagrammed in the figure 5.5.

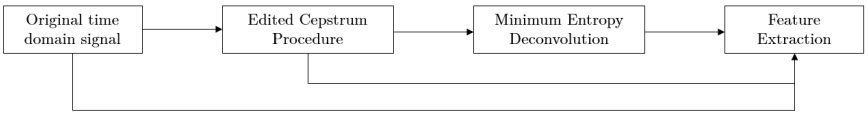


Figure 5.5: Diagram of the procedure for feature extraction

5.2.2 Features Correlation

After observing the variables evolution with speed and load it was thought, as an interesting point, to evaluate the correlation coefficients between the features. The goal is understand which features are related between them to use a minimum of features that allow an effective fault detection in rolling element bearings. The correlation coefficient of two random variables is a measure of their linear dependence [46]. For a variable with length N, which in our case will be the number of speed measures, the Pearson correlation coefficient is defined by equation (5.16). Where μ_A, μ_B and σ_A, σ_B are the mean and standard deviation of A and B respectively. Equation (5.17) define the correlation coefficients in terms of the covariance of A and B.

$$\rho(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\overline{\frac{A_i-\mu_A}{\sigma_A}}\right)\left(\frac{B_i-\mu_B}{\sigma_B}\right) \tag{5.16}$$

$$\rho(A,B) = \frac{cov(A,B)}{\sigma_A\sigma_B} \tag{5.17}$$

The correlation coefficient matrix of two variables is the matrix of correlation coefficient for each pairwise variable combination, as seen in equation (5.18). The unit value in the matrix diagonal are due to the fact that the variables correlate to themselves in a perfect way which rises a correlation coefficient of 1.

$$R = \begin{bmatrix} 1 & \rho(A,B) \\ \rho(B,A) & 1 \end{bmatrix} \tag{5.18}$$

To easily identify the values, in the Chapter 6, the results are presented in a colour matrix as exemplified by figure 5.6, where the axis represent the features in study. The upper triangular matrix represent one load scenario and the lower triangular matrix another load case when applied. In cases where only on load was use, the matrix is symmetrical since $\rho(A, B) = \rho(B, A)$. The goal of feature trending is the identification of patterns by using a minimum set of uncorrelated features. In the next section, a method will be discussed, that is focus on the most relevant features which are used as the basis to develop a maintenance strategy.

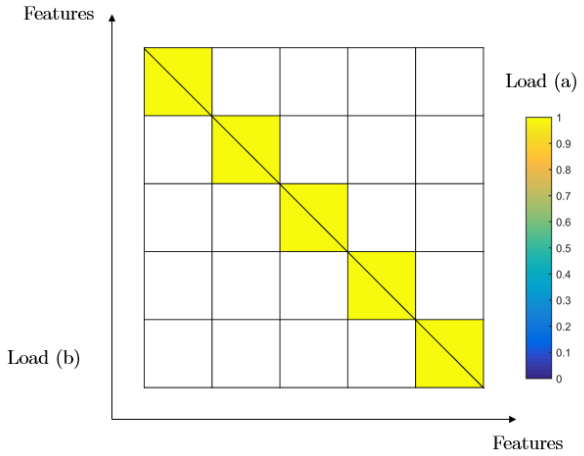


Figure 5.6: Correlation coefficient matrix used to present the result

5.2.3 Principal Component Analysis

The Principal Component Analysis (PCA) is a way of identifying patterns in data and expressing that data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find for high dimension, this technique presents a powerful tool for data analysing [47]. Usually, this approach can only be effectively performed on a set of observations which vary linearly [48]. A new approach called Kernel Principal Component Analysis was studied by [48] which suits the non-linearities of the highly stochastic data, providing from operational and changing conditions on slow speed bearings. However, our main interest in this section is to explain the way to apply this method in the features obtained from the vibration signals.

PCA is a projection statistical method used for dimensionality reduction. It produces a lower dimensional representation in a way that preserves the correlation structure between the variables [49]. Given a set of m variables with length n stacked into a matrix X whose covariance matrix has eigenvalues λ and eigenvectors p such that equation (5.19) verifies.

$$(\lambda_1, p_1), (\lambda_2, p_2), \cdots, (\lambda_m, p_m) \qquad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_m \geq 0 \tag{5.19}$$

The principal decomposition component of X can be represented by equation (5.20), where $T = [t_1, t_2, \cdots, t_l]$ is defined to be the matrix of principal component scores, $P =$

$[p_1, p_2, \cdots, p_l]$ is the matrix of principal component loadings, E is the residual matrix in the sense of the minimum Euclidean norm and l is the index of the principal components [49].

$$X = TP^T + E \ = \ \sum_{i=1}^{l} t_i p_i^T + E \tag{5.20}$$

The identification of the PCA model consists in estimating its parameter by an eigenvalue-eigenvector decomposition and determining the number of components l to retain. According to [49], many procedures have been proposed for selecting the number of components to retain. One approach is the experimental method which judges the cumulative sum contribution of the anterior l component is higher than 0.85 as stated by equation (5.21).

$$100 \cdot \frac{\sum_{i=1}^{l} \lambda_i}{\sum_{i=1}^{m} \lambda_i} > 85\% \tag{5.21}$$

An easy algorithm to perform the technique and obtain the new data is detailed below.

1. Evaluate the mean of each variable and subtract that value from each of the data dimension. The mean subtracted is the average across each dimension and will produce a data set with zero mean.

2. Obtain te covariance matrix which will be a square matrix with dimension equal to the number of variables. The covariance is expressed by the formula. The covariance matrix of two variables is the pair wise covariance between each variable.

$$cov(A, B) = \frac{1}{N-1} \sum_{i=1}^{N} (A_i - \mu_A) \times (B_i - \mu_B) \tag{5.22}$$

3. Calculate the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors should be unitary. If justified, plot the eigenvectors above the original data. These vectors provide information about the patterns in the data. The next steps will express these data in terms of the eigenvectors.

4. Organize the eigenvalues from highest to lowest which will give the components in order of significance. Instead of using equation (5.21) to choose the components to keep, the analyst can look to the eigenvalue and assess the importance of the components.

5. Apply equation (5.20) to obtain the decomposition of X.

The final step will give the original data in terms of the vectors chosen. The original data was set in terms of the model variables. If the axes are perpendicular, the expression is the most efficient, this is why it is important that the eigenvectors are perpendicular to each other [47]. The kernel principal component analysis is similar to this procedure but used the kernel matrix instead of the covariance matrix and also a centring operation for the features [48].

Chapter 6

Experimental Tests and Results

The results presented in this section were obtained through two sets of measurements. One of them in the facilities of the Flander Mechatronics Technology Centre (FMTC), recently renamed Flanders Make which is a research centre for the manufacturing industry, located in Leuven (Belgium) and that collaborates with research laboratories at 5 Flemish universities. This centre allocate considerable part of their efforts to international cooperation in the field of innovation and take part into several European research projects. Their goal is to have more than 300 researchers working full time, creating a unique community on industrial research by the year of 2018. The second part of the tests were conducted at the facilities of the company were this text was developed, also in Leuven. Siemens SISW is a world-leading provider of product life cycle management and manufacturing operations management software. The goal of the company is to improve and optimize their clients processes through a properly planning and development of manufacturing and support. The company has around 77 000 clients and 1200 employees worldwide and is responsible for more than 40% of the 3D data generated globally. This second set of test were conducted in a machine fault simulator provided by SpectraQuest (SQ) and the regime of tests includes both low and medium regime speeds while the test conducted on FMTC were for low speed regime exclusively.

6.1 Flanders Make

6.1.1 Operational Conditions

The measurements conducted in the FMTC facilities covered the low-speed range. The test rig used was the one showed in figure 6.1 which consists of an electric motor, controlled by a variable frequency drive, connected to a gearbox through a flexible coupling. The input shaft of the gearbox was set to run at 180, 480, 1200 and 2400 RPM. The bearing under analysis was located on the intermediate shaft of the gearbox which has a gear ratio of 29/100 which means that the rotational speeds of the studied bearing shaft are 52, 139, 348 and 696 RPM composing 4 cases. At the output shaft it is coupled a magnetic-particle brake with a maximum torque of 27.5 Nm which was adjusted to 25% (6.9 Nm) and 100% (27.5 Nm) which will be designated Load 25 and Load 100 respectively. For the low speed regime measurements the sensors used were 1 acoustic emission, 1 accelerometer high-end in the y direction, 1 accelerometer high-end in the direction x and y and 1 DC MEMS accelerometer. The procedure which was followed for assembling the bearings is detailed in appendix B.

Table 6.1: Tested bearing geometric properties

ER-16K	
Ball Diameter	7.94 mm
Pitch Diameter	39.32 mm
Number of Balls	9
Contact Angle	0^0

The bearing conditions studied were the Healthy and Inner race fault cases and the bearings used were ER-16K which geometric properties are declared in the table 6.1. Due to lack of time, it was not possible to correctly measure the 139 RPM with the Load 100 for the outer race fault case and therefore this case will not be contemplated.

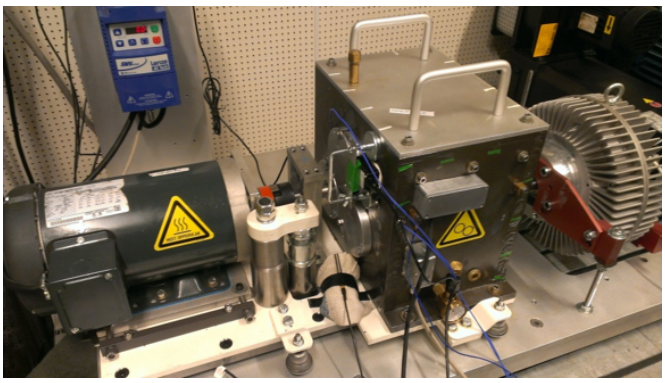


Figure 6.1: Fault simulator from FMTC

For the test bench in the Flanders Make facilities the accelerometer which returned the best results was the one located close to the housing where was located the defected bearing. This section will be divided in two parts. The first one, relative to the envelope analysis and the second one to the features extraction. After processing all the data recorded it was chosen the speed of 139 RPM and y direction, explained in figure 6.2, as the one that returned more interesting and accurate results.

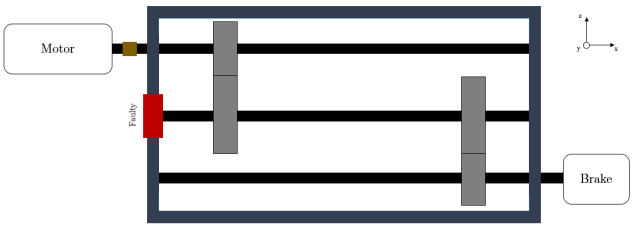


Figure 6.2: FMTC Envelope - Axis definition

6.1.2 Envelope Analysis

6.1.2.1 Race Faults

The results, using the envelope algorithm adopted in this text, for the inner race fault in the FMTC test bench with a Load of 27.5 Nm, input speed of 1200 RPM and shaft speed of 139 RPM is presented below, in the figure 6.3. The parameters used for the algorithm and other relevant information is presented in the table 6.2.

Table 6.2: FMTC Envelope - Inner race fault parameters

General		**MED**	
BPFI	31.22 Hz	Filter Size	120
ω_{Mean}	19.9 Hz	Δ_{Iter}	1
Discrete Removal		**Fast Kurtogram**	
Nrahm	120	Central Freq.	12 kHz
r	0.15	Bandwidth	24 kHz

Due to time constrains it was not possible to get accurate results for the outer race fault. The test bench for the low speed rotation is shared with other entities and it was not possible, during the time available for this study measurements, to detect and solve the problem. Further investigation is needed at the FMTC test bench in order to achieve valid results and this is the reason why the outer race fault condition is not presented in this chapter.

Inner Race Processing

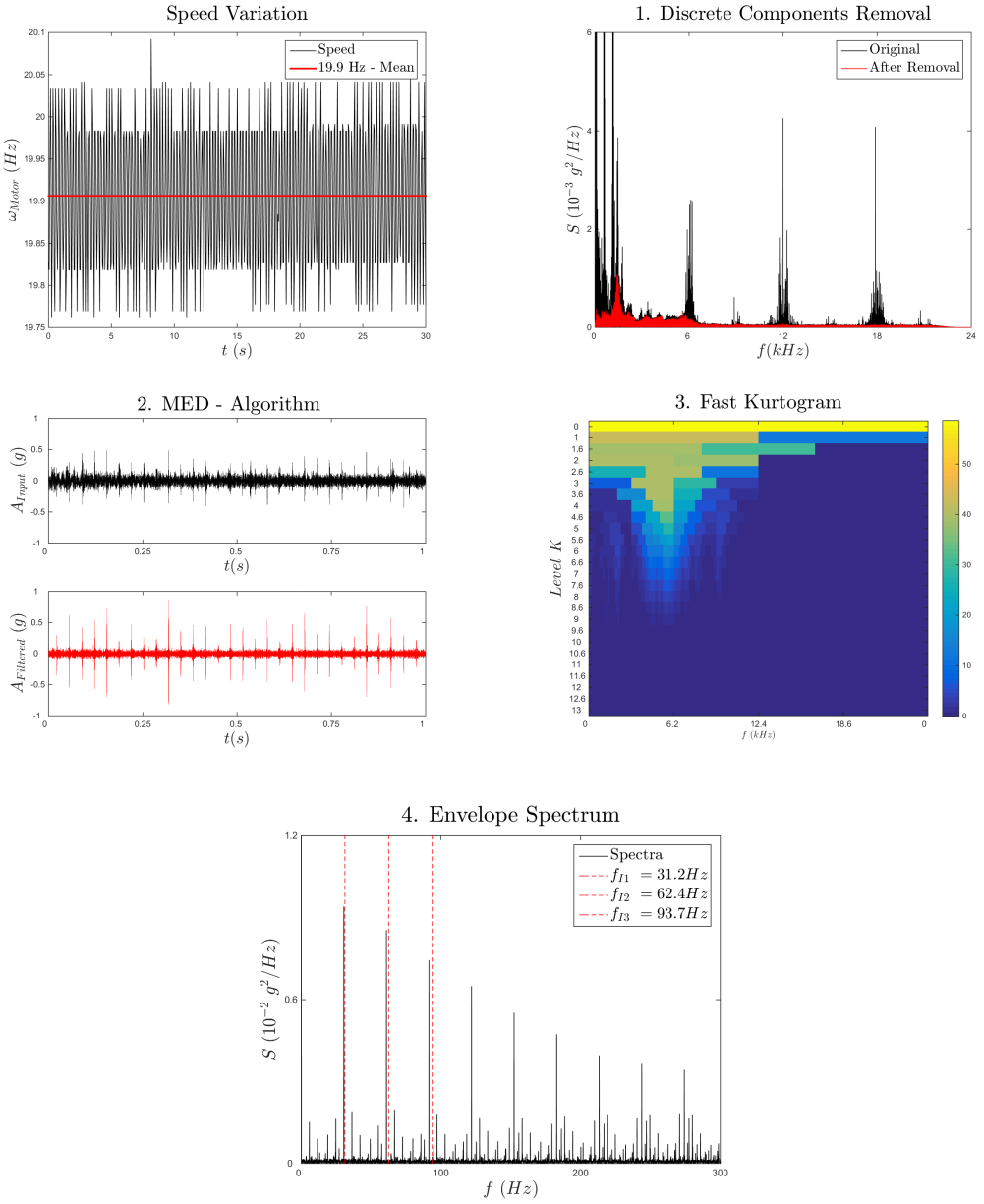


Figure 6.3: FMTC Process - Inner race fault at 139 RPM

Comment 1. Figure 6.3 shows the results for the inner race fault. The first step shows the removal of the discrete components which for the FMTC test bench were due not only to the motor rotation but also the gearmesh frequencies. In the second step, representing 1 second of the measured signal, we can clearly notice the MED algorithm performing its task and the enhancement of the pulses in the signal. The fast kurtogram reaffirms that the overall signal is very impulsive and because the maximum spectral kurtosis occurred in the level 0 the signal will be demodulated in all its band which means that no demodulation occurs and so the envelope analysed is the envelope of the full signal. For last, the BPFI and its harmonics are clearly prominent in the spectrum allowing an easy detection of the fault.

6.1.2.2 Healthy Comparison

In this section are presented the respective spectrum of the inner and outer race faulty bearings and the healthy bearing running in the same conditions.

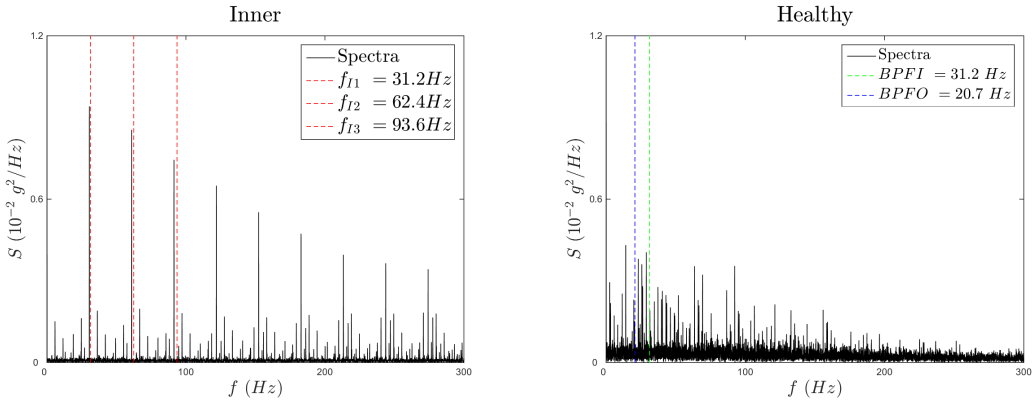


Figure 6.4: FMTC - Healthy and Faulty comparison

Comment 2. As had been said, the fault is detected for the inner race fault, however, it is interesting to compare the results with the spectrum of an healthy bearings under the same operational conditions. As we can observe, the magnitude of the spectrum for the fault frequency in the inner race fault case is almost twice the average magnitude for the healthy case one. This fault is more difficult to detect than the outer race fault, because the race is continuously rotating and the fault may not be under load when a rolling element strikes it.

6.1.2.3 MED Influence

The application of the MED algorithm can be evaluated on the following figures representing the inner and outer race faulty bearings with and without the MED algorithm applied.

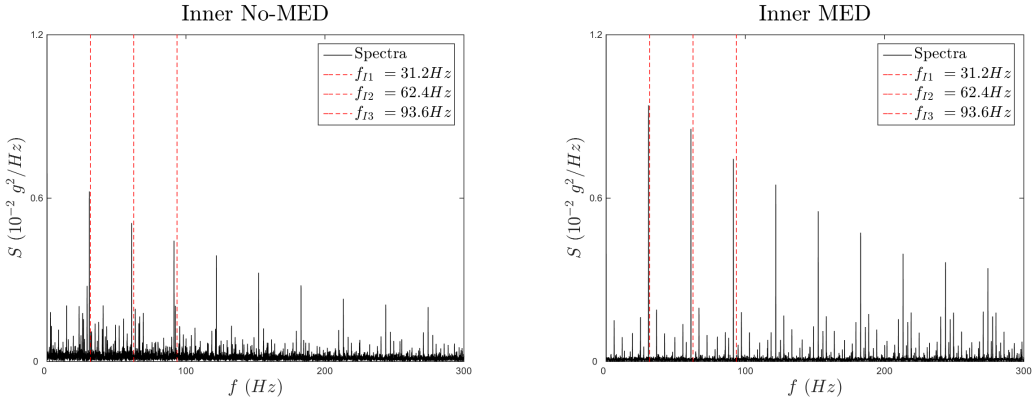


Figure 6.5: FMTC - MED algorithm influence evaluation

Comment 3. The influence of the MED algorithm is clear in the envelope analysis, enhancing the impulsiveness of the signal and allowing an easier detection of the faults. The

magnitude of the fault frequency increased around 50% after the algorithm application for the inner race fault. Despite this, in some cases the algorithm may return worse results due to interferences with highly impulsive nature.

6.1.3 Features Extraction

The features enumerated in the Chapter 5 were applied to the range of speeds and loads evaluated. In this section, it was chosen to eliminate the speed of 52 RPM which reduces the data points to 3 speed operational ones.

6.1.3.1 Evolution with Speed

In here, it will be presented the evolution of the feature Kurtosis, RMS and Peak-to-Peak for the inner race fault with the rotational speed. On the left is shown the Load 25 and on the right the Load 100.

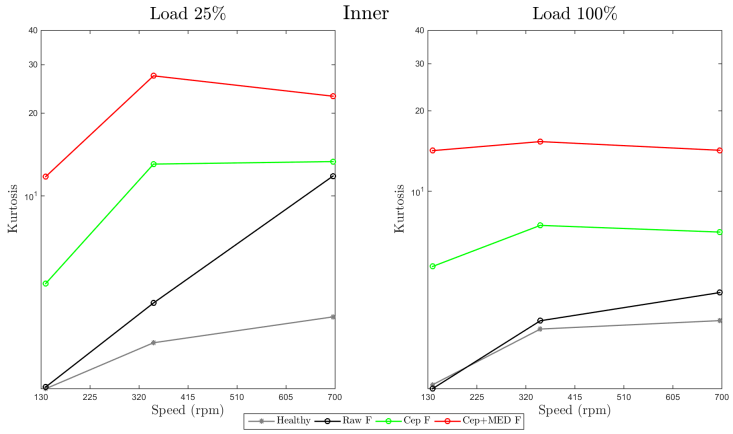


Figure 6.6: FMTC Features - Evolution of the kurtosis with speed

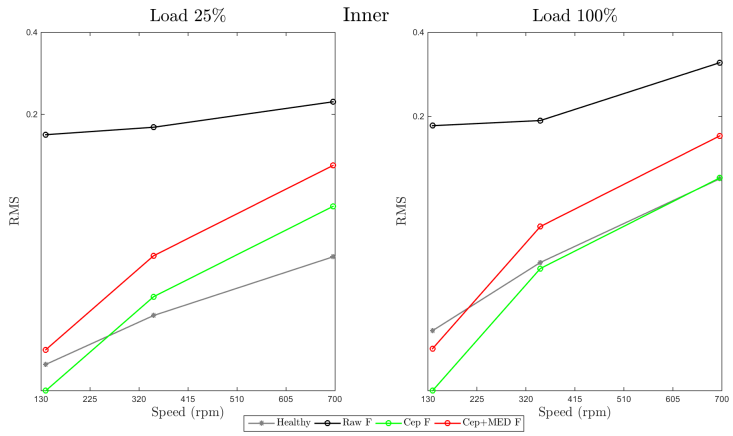


Figure 6.7: FMTC Features - Evolution of the RMS with speed

Comment 4. Figure 6.6 represent the difficulty to evaluate feature trending in low speed rotations. It is observable that when the speed increases, the difference between the healthy and faulty case is more notorious. In this case, the application of the procedure used for envelope analysis improved the detection of the fault, in both cases, applying only the discrete components removal and even better with the MED. However, because these techniques change the nature of the time signal, an effective detection method can not rely on them. Figure 6.7 shows a clear distinction between the faulty and healthy cases, however, when the discrete components are removed and the MED is applied the signal, despite keeping the trending, there is an intersection with the healthy curve. Therefore, it is not possible to determine whether the procedure adopted for envelope analysis improve the fault detection based on feature extraction.

6.1.3.2 Filter Influence

The data obtained in the facilities of the FMTC were processed twice, with different filter parameters in the MED algorithm. Figure 6.8 shows the filter size L=120 with stopping criteria equals to 1 and the filter size L=120 but with the stopping criteria 2.5.

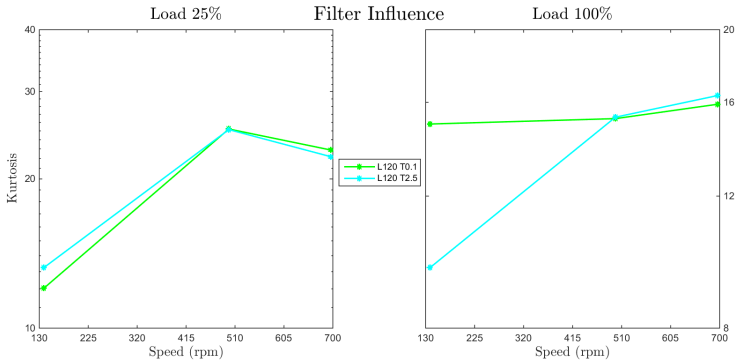


Figure 6.8: FMTC Features - Influence of the filter parameters

Comment 5. The processing with different parameters applied on the MED algorithm showed that the stopping criteria does not considerably change the final results. Despite the apparently change for the Load 100%, the analysis of other features allowed the same conclusion obtained for the Load 25% on the left side of figure 6.8.

6.1.3.3 Autocorrelation Matrix

The autocorrelation matrix allows the evaluation of the relation between features for a possible combinations of several of them. In the next figure are presented the results for the inner race fault regarding the correlation matrix, with the evolution of the procedure implemented for envelope analysis. The upper triangular part refers to the Load 100% and the lower part to the Load 25%.

Comment 6. The correlation matrix for the unprocessed data does not allow any conclusions. After the removal of the discrete components it is possible to perceive some changes that become clear after the MED algorithm. Very high correlations are produced between the RMS, the percentile 75, the Shannon Entropy and the Log entropy. On the other hand features

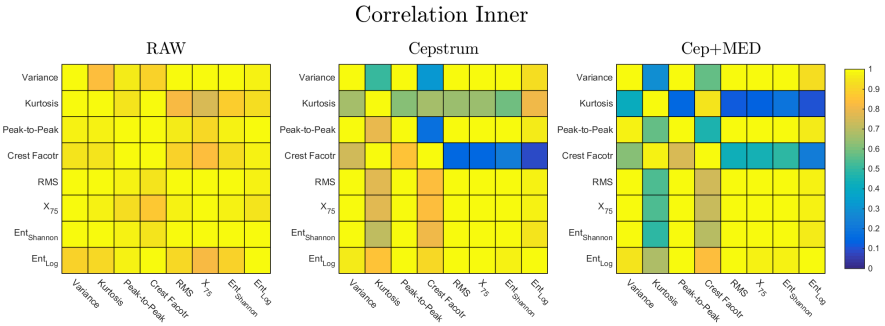


Figure 6.9: FMTC Features - Autocorrelation matrix for the inner race fault

like variance, crest factor and peak-to-peak seem not to correlate so well and should probably be used as single indicators.

6.2 SpectraQuest

6.2.1 Operational Conditions

The machine fault simulator from SQ is the one shown in figure 6.10. This equipment allows to mount two bearings on a shaft, powered by an electric motor controlled by a simple variable frequency drive.

Table 6.3: SQ operational conditions

Cases	Healthy, Inner and Outer
Loads	LM and LR
Speeds	300:120:2700

The engine was a tree-phased electric motor of 0.5 HP with nominal speed of 3450 RPM and rated frequency of 60 Hz. As clarified in table 6.3 the range of speeds measured was from 300 up to 2700 rpm with increments of 120 rpm totalizing 21 cases, this means that part of the low speed regime is covered and also a large part of the medium regime. In addition to this, a static load of 5 kg was added to the shaft to increase the amplitude of the faults in the bearings. The load was placed in the middle of the shaft, refereed from now on as Load Middle (LM) and at 3 cm from the bearing under analysis, which in the configuration adopted was in the right of the shaft and will be from now on defined as Load Right (LR). The cases studied in these tests were the Healthy, Inner and Outer race ones. The bearings were placed in the right housing while the left housing kept a healthy bearing for all the measurements. The bearings used were ER-16K whose geometric properties are detailed in the table 6.1. For the SQ machine, the sensors used were 2 triaxial accelerometer, 2 acoustic emission sensors and 2 microphones placed in the healthy, to record the validation results, and in the faulty side.

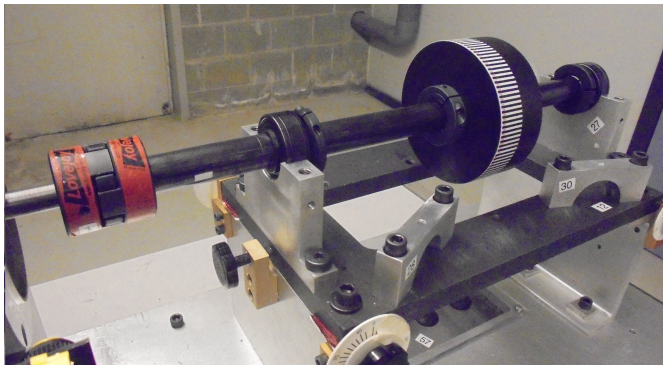


Figure 6.10: Fault machine simulator from SpectraQuest

For the test bench in the Siemens PLM facilities the accelerometer which returned the best results was the one located close to the housing where was located the defected bearing. This section will be divided in two parts. The first one, relative to the envelope analysis and the second one to the features extraction. After processing all the data recorded it was chosen the speed of 1500 RPM and y direction, explained in figure 6.2, as the one that returned more

interesting and accurate results.

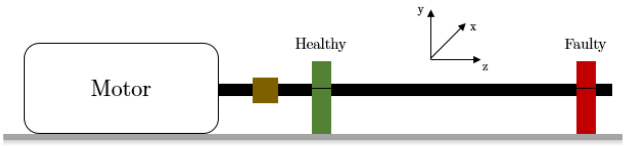


Figure 6.11: SQ Envelope - Axis definition

6.2.2 Envelope Analysis

6.2.2.1 Race Faults

The results, using the envelope algorithm adopted in this text, for the inner and outer race fault in the SQ test bench with a load placed in the middle of the shaft, with a shaft speed of 1500 RPM are presented below, in the figures 6.12 and 6.13 respectively. The parameters used for the algorithm and other relevant information are presented, in the same way, at the tables 6.4 and 6.5.

Table 6.4: SQ Envelope - Inner race fault parameters

General		**MED**	
BPFI	136.78 Hz	Filter Size	120
ω_{Mean}	1517.4 rpm	Δ_{Iter}	1
Discrete Removal		**Fast Kurtogram**	
Nrahm	∞	Central Freq.	6.4 kHz
r	0.15	Bandwidth	12.8 kHz

Table 6.5: SQ Envelope - Outer race fault parameters

General		**MED**	
BPFO	90.91 Hz	Filter Size	120
ω_{Mean}	1518.7 rpm	Δ_{Iter}	1
Discrete Removal		**Fast Kurtogram**	
Nrahm	∞	Central Freq.	12.5 kHz
r	0.15	Bandwidth	533.3 Hz

Inner Race Processing

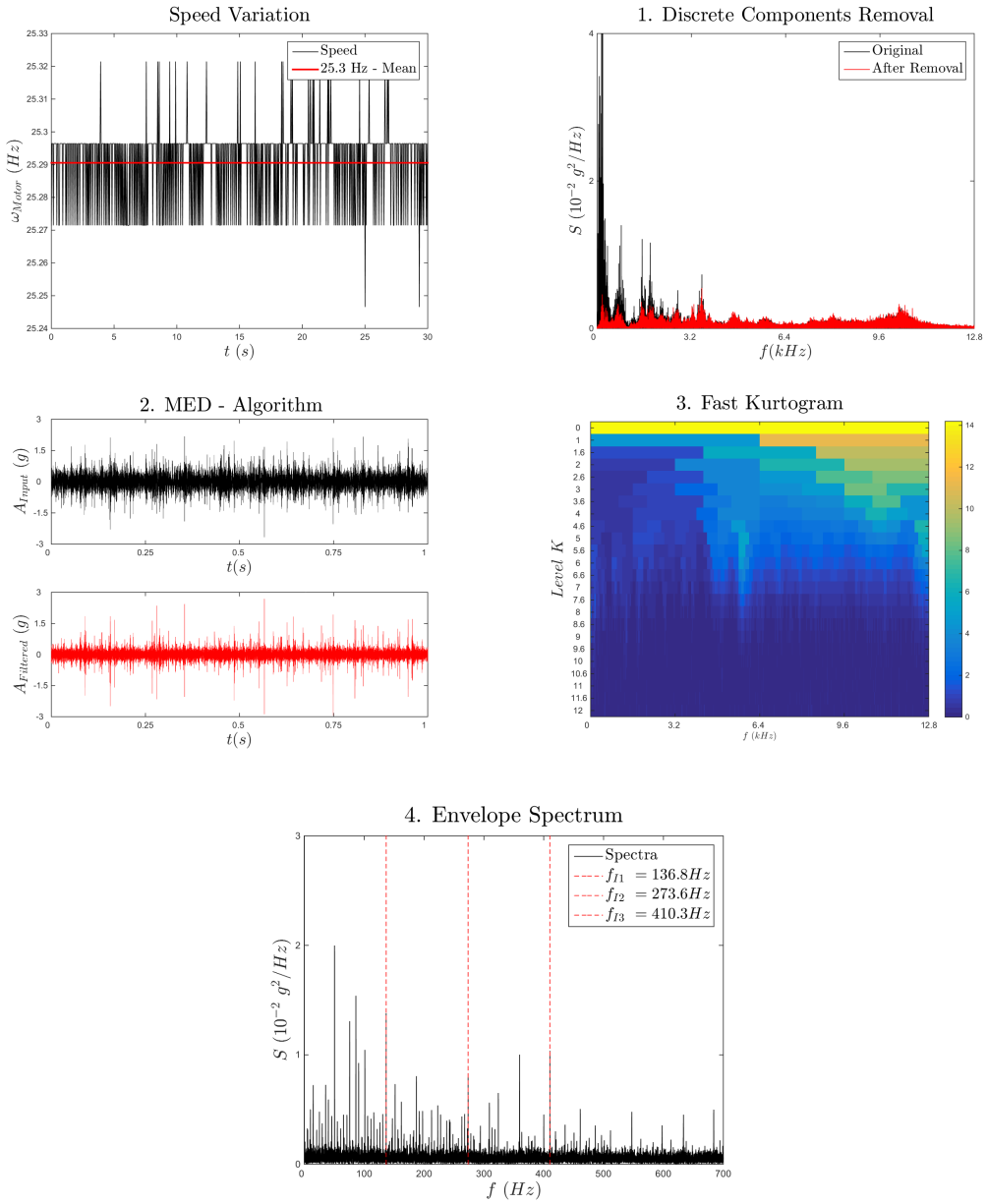


Figure 6.12: SQ Process - Inner race fault at 1500 RPM

Comment 1. For the inner race faulty bearing tested on the SQ, the components removed on the first step are only originated from the shaft rotation and the MED enhances the impulsiveness of the signal by maximizing the spectral kurtosis. Despite the impulsiveness noted on the 1 second representation before and after the MED application, and also the fast kurtogram, which returns the signal as impulsive through all its bandwidth, the fault is not easily detected. Although the components in the spectrum, related to the BPFI are notorious, the sidebands magnitudes are high compared to the fault, which is smearing the results.

Outer Race Processing

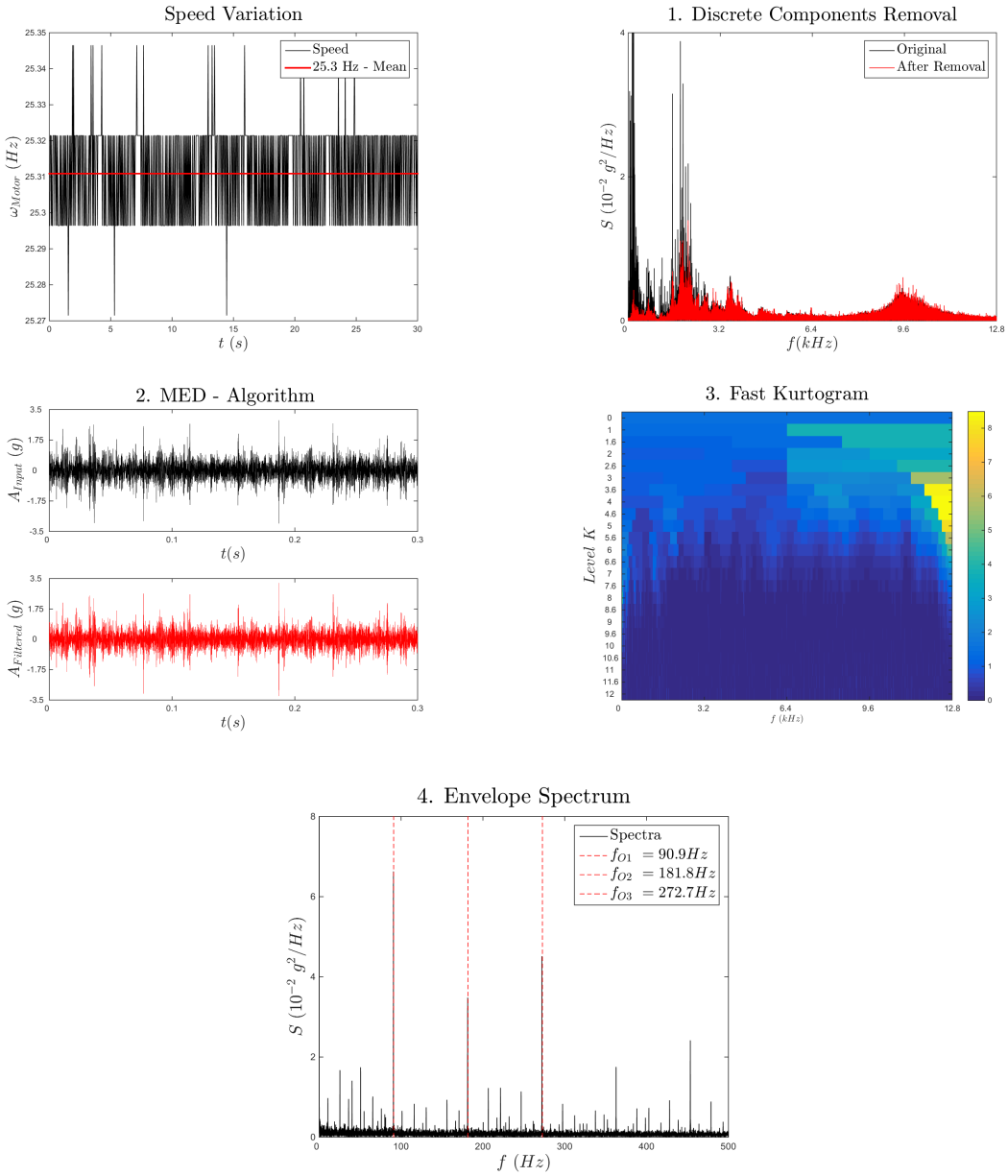


Figure 6.13: SQ Process - Outer race fault at 1500 RPM

Comment 2. Contrary to what happened for the FMTC case, here the better results were achieved for the inner race fault. In this case, with the removal of strong periodic components as seen before and after the discrete components removal, the envelope of the narrow band were the signal was demodulated, enable a clear detection of the outer race fault, as seen in the envelope spectrum. The BPFO frequency and the first two harmonics are clearly highlighted from the rest of the spectrum.

6.2.2.2 Healthy Comparison

In this section are presented the respective spectrum of the inner and outer race faulty bearings and the healthy bearing running in the same conditions.

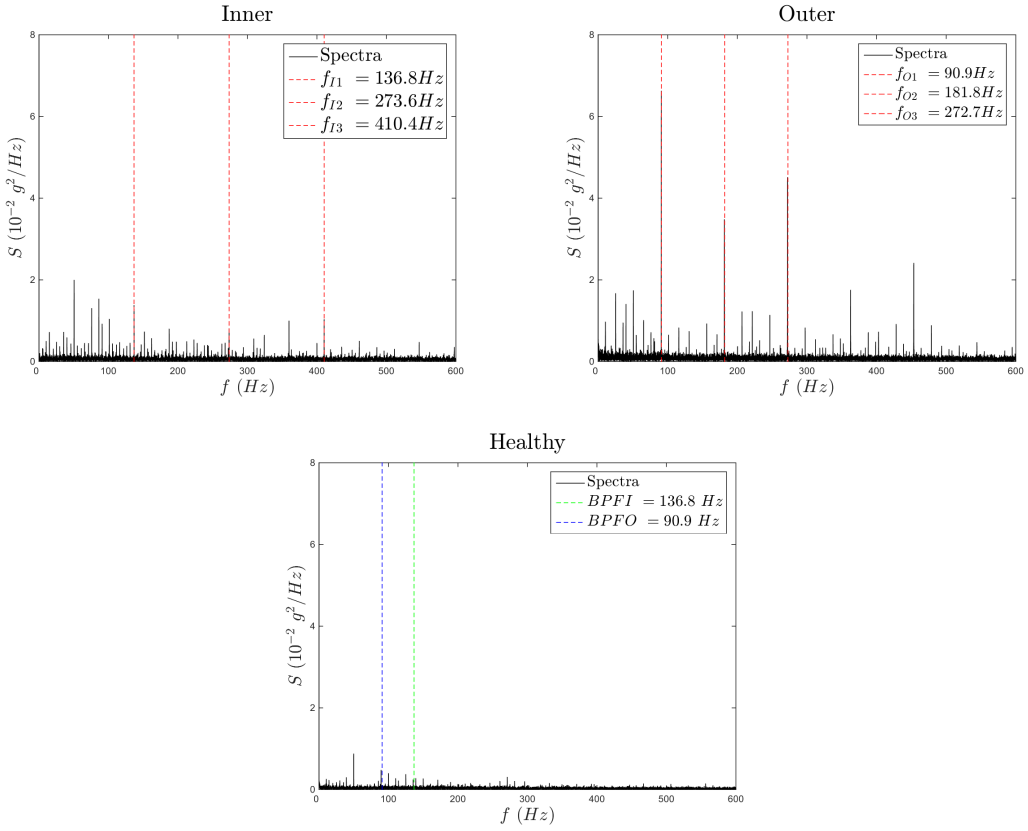


Figure 6.14: SQ - Healthy and Faulty comparison

Comment 3. Regardless the credibility of the inner race results, the magnitude of the fault frequency in the inner case is clearly higher than for the healthy one. The outer race fault frequency shows a magnitude more than 5 times higher in the faulty case than in the healthy one. These results present what is expected from a faulty bearing which is higher vibration levels for the FCF which would ring the alarms in a CBMS.

6.2.2.3 MED Influence

The application of the MED algorithm can be evaluated on the following figures representing the inner and outer race faulty bearings with and without the MED algorithm applied.

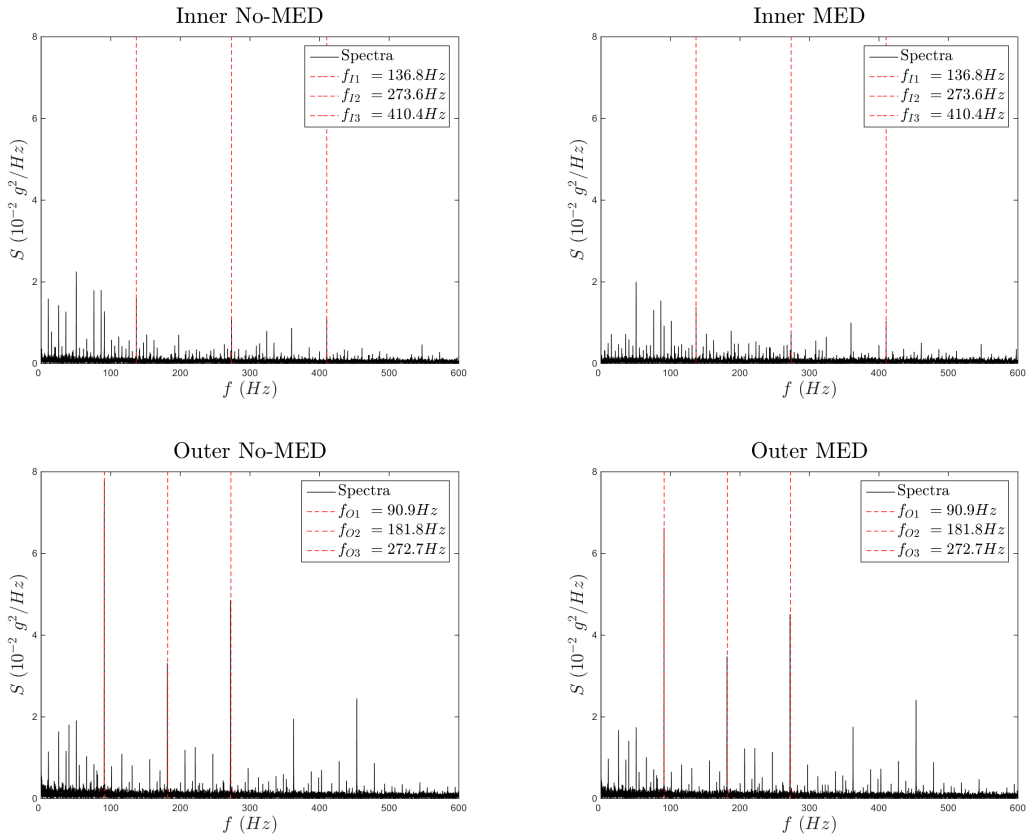


Figure 6.16: SQ - MED algorithm influence evaluation

Comment 4. For the inner race fault, the MED algorithm does not improve the fault detection. In fact, if the signal as a lot of interferences, the algorithm will not perform properly, trying to enhance spikes that may not be related to the fault and could even worsen the results. As seen for the outer race fault, the MED algorithm reduced the magnitude of the fault frequency and due to this fact is advisable to plot the results with and without the application of the MED algorithm.

6.2.3 Features Extraction

The features enumerated in the Chapter 5 were applied to the range of speeds and loads evaluated. In this section, it was chosen to eliminate some of the speed points due to their inaccuracy which left the results in the range between 780 and 2340 RPM which reduces the data points to 14 operational speeds.

6.2.3.1 Evolution with Speed

In here, it will be presented the evolution of the features Kurtosis and RMS for the inner and outer race fault with the rotational speed. On the left is shown the Load Middle and on the right the Load Right.

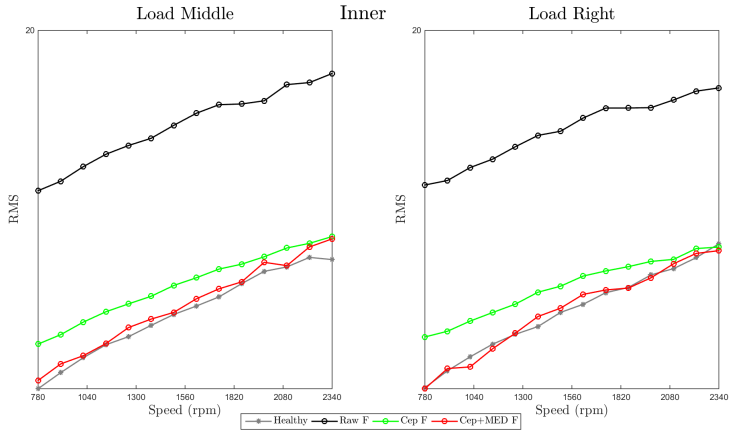


Figure 6.17: SQ Features - Evolution of the RMS for the inner race fault

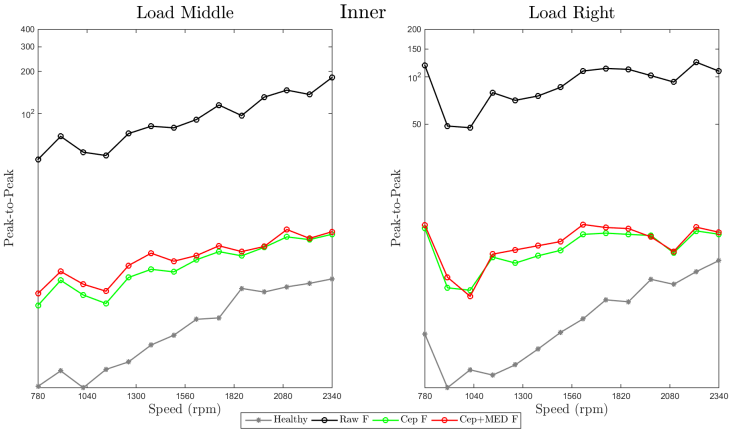


Figure 6.18: SQ Features - Evolution of the Peak-to-Peak for the inner race fault

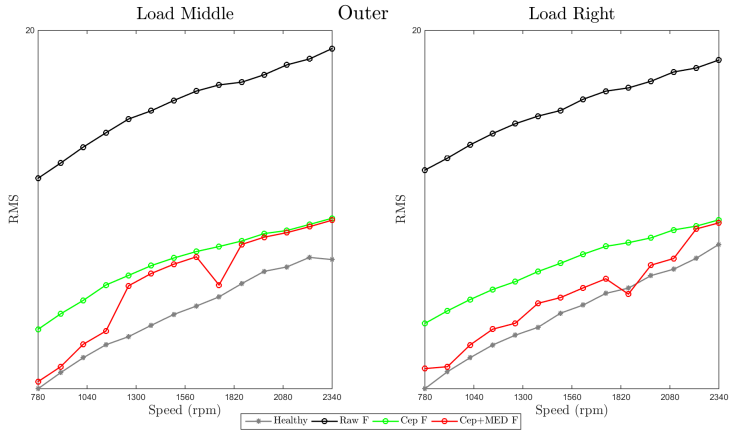


Figure 6.19: SQ Features - Evolution of the RMS for the outer race fault

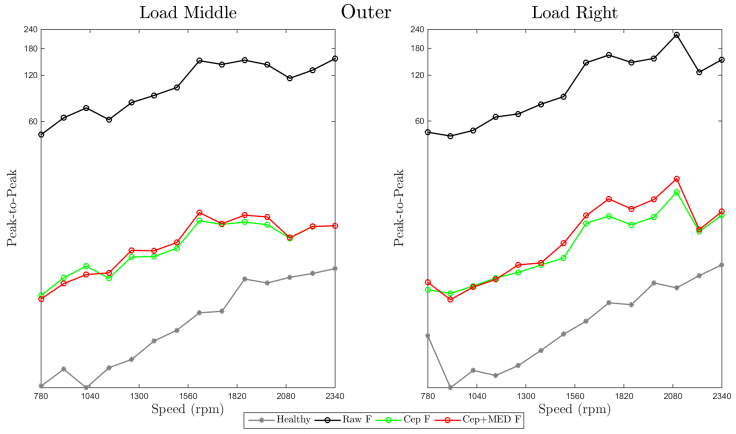


Figure 6.20: SQ Features - Evolution of the Peak-to-Peak for the outer race fault

Comment 5. For the medium speed regime, is unanimous that the procedure used for the envelope analysis does not improve the fault detection. In fact, the cepstrum editing is the best procedure to adopt. The grey curve, representing the healthy case had the discrete components removed, as well as the green curve for the faulty case. The trending for the inner race on the RMS is clear and the faulty levels much higher than the healthy ones. Applying the MED creates non-conformities in the trend of RMS for the outer race fault however there is a distinct separation between faulty and healthy. For the peak-to-peak, the results, while returning what was expected, exhibit some fluctuations which allows to state that this is a sensitive parameter. This is due to the fact that it measures the difference between the highest and lowest value and is likely to be affected by random interferences.

6.2.3.2 Filter Influence

The data obtained in the facilities of Siemens PLM were processed with different filter parameters in the MED algorithm. Figures 6.21 and 6.22 show the different configurations with 6 different filter sizes, $L = 60, 120, 200, 275, 350, 500$, and 4 different stopping criteria $\Delta_{Iter} = 0.1, 0.8, 1.5$ for the inner and outer race faults with Load 25. Each one of the 24 configurations took around 1,5 days to process all the data.

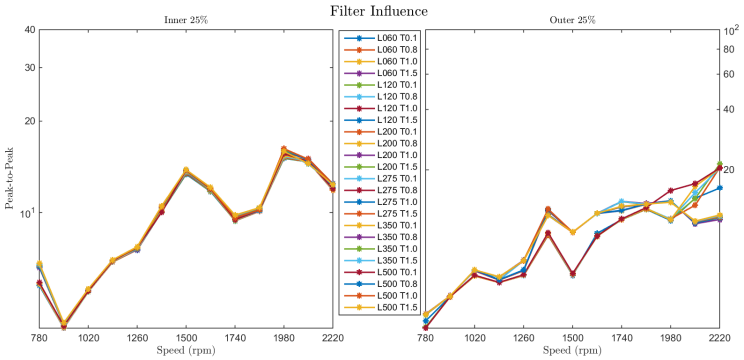


Figure 6.21: SQ Features - Influence of the filter parameters on the Peak-to-Peak

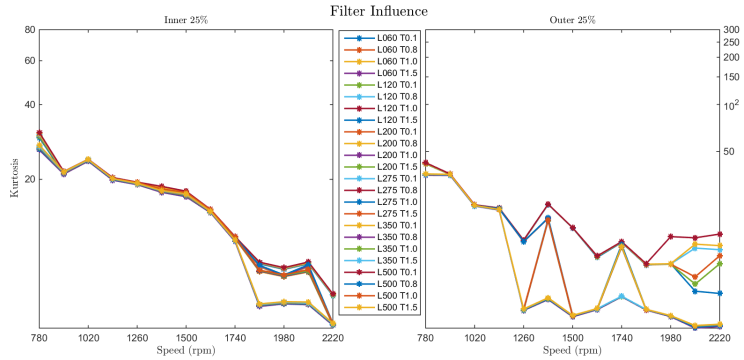


Figure 6.22: SQ Features - Influence of the filter parameters on the kurtosis

Comment 6. The parameters used for the MED algorithm do not influence considerably the final results. In fact, for the inner race fault all the different cases are coincident and for the outer race fault a slight dispersion is observed. Despite being sensitive, the peak-to-peak present better robustness than the kurtosis as regards to the filter parameter influence.

6.2.3.3 Autocorrelation Matrix

The autocorrelation matrix allows the evaluation of the relation between features for a possible combinations of several of them. In the next figures are presented the results for the inner and outer race fault cases regarding the correlation matrix, with the evolution of the procedure implemented for envelope analysis. The upper triangular part refers to the Load Right and the lower part to the Load Middle.

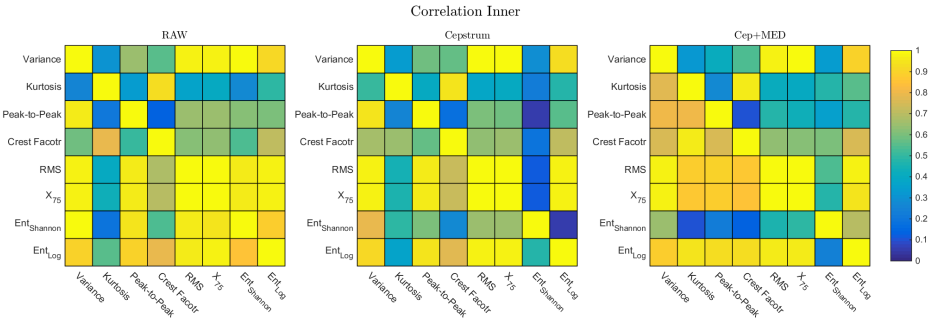


Figure 6.23: SQ Features - Autocorrelation matrix for the inner race fault

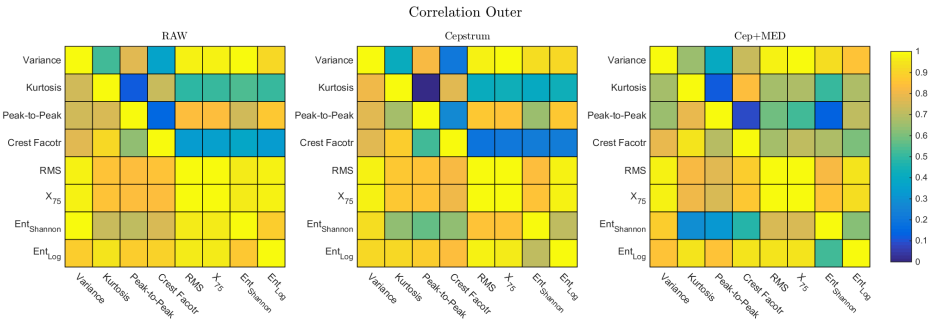


Figure 6.24: SQ Features - Autocorrelation matrix for the outer race fault

Comment 7. For the inner race is notorious some correlation between the RMS, percentile 75, variance and kurtosis when the load is located in the middle, that decreases considerably when the load is located on the right. This contradicts what happened for the low speed regime. On the outer race fault, there is a good correlation between almost all the features plotted, when the load is located on the middle and once again the value decrease when the load is located near the fault.

Chapter 7

Conclusion

7.1 Conclusions

After all the research and tests conducted during this project several conclusions were formed. The vibration monitoring is the most widely applied technique for condition monitoring with accelerometers returning good detection of the fault, while some authors also proposing acoustic emission sensors to detect incipient faults in the early stages, what would provide more time to schedule maintenance. One of the most used technique in vibration monitoring is the envelope analysis which allows the analyst to detect the presence of the faulty frequencies. Other approaches such as the wavelet transform are also being heavily studied with favourable results that potentiate its applicability. The removal of the periodic components or discrete components allows better detection due to the filtering that is conducted on the signal. For impulsive signals, when there are low interferences, the MED algorithm improve considerably the results and the magnitude of the fault frequencies by enhancing the signal impulsiveness. However, it is advisable to always represent the final results with and without the application of this method due to the fact that in some cases, where the external interferences present also high impulsiveness, the method could converge to a solution which is not the desired fault and the magnitude of the last would be affected.

From the features extraction point of view it can be stated that only a few features from all the ones studied, returned accurate results and therefore, a maintenance strategy based only on the feature trending has to be carefully planned not to give false alarms or on the other hand miss the fault detection. The more interesting parameters and because of this, plotted in the results chapter, were the kurtosis, the RMS and the peak-to-peak. However, other parameters such as the variance and the Shannon entropy returned interesting trending. The accuracy of each feature as to be assessed for each application and that is the reason why a strategy based in several parameters, acting as single individuals or composing a new parameter is an interesting approach for a future work. This text showed the correlation between some of the features studied and good relations were found between the RMS, the percentile 75, the Shannon entropy for the low speed regimes and between the kurtosis, variance and RMS for the medium speed regime. The adopted method used for envelope analysis was used in combination with the features extraction, however its relevance is doubt, while in some cases it improved the difference between the healthy and faulty curves in others, it brought the faulty curve close to the healthy one, not allowing the fault detection. Because it is still in a developing phase, the feature extraction approach should be used with caution with some work needed to be done in order to make possible an effective clean of the signals before trending, by properly designed filters and also an improve in the aggregation with other existing techniques.

7.2 Future Developments

Despite all the techniques developed until now, there are several aspects that need to be further investigated to allow automated detection systems to be used in an efficient and reliable way. According to Heng et al. [50] there are three main aspects that need attention at this moment.

1. Methods which allow the use of available data in a more fully and accurately way. Also the combination of different techniques, as well as the range of machine types, needs to be more flexible.

2. Develop ways to consider the complexities of variable operating conditions which occur when machines are subjected to changes in the operating parameters or repairs. This changes can reduce the accuracy of the prognosis output.

3. Due to the equipments complexity, it is needed a system which has into account different failures modes of the individual components as well as the interaction between these components with the operating environment.

The analyst of a group of identical equipments must have in consideration that each unit is an individual system and the information of one system does not reflect the population characteristics. In addition to this, the failure behaviour of each unit is a function of the work schedule, the environment and the duty parameters. The data provided by CM is normally good for short time condition prediction and the predictions made by the current prognostic models perform single steps to estimate the vibration levels which will be present in the next immediate time step. Indeed, it is needed to develop techniques which allow sufficiently long lead time to properly prepare the next maintenance. Or even, a longer prediction could be used to evaluate the viability of keeping the equipment running until the next maintenance opportunity, for instance after the end of a batch or until the next scheduled inspection [50]. These predictions could be used in combination with the remaining useful life information, and have into account historical data such as the expected life of a component that had the same operating conditions in other equipment and not only rely on the manufacturer information.

An addition to what was said in the previous paragraph is the problem of the censored data. This concept is present when the units are replaced when a failure is detected and it is common in applications were downtime is not admissible such as medical or military applications. This means that the units replaced do not really fail, thus the point at which the components would actually fail is not known or recorded. Usually, the point recorded as fail is the point until which the equipment had survived without failure and it is not difficult to find components that will run for as much time with a failure such as without one. In other words, the remaining useful life is sometimes similar to the healthy life. This means that create models to predict other components failure based on censored data means that an underestimation may occur. The analyst must recommend a maintenance schedule that does not interrupt production or wastefully replace units that still have useful remaining life and so a correct evaluation of the underestimation error is needed. On the other hand, an easy temptation would be to omit censored data, however this would mean a reduction on the amount of available information which sometimes takes years to get [50].

When a component is repaired in a machine it does not mean that the condition of the overall equipment is reset to as good as new and therefore, some effort needs to be spent into prognostics models which account for the effects of maintenance actions. The optimization in this field can be achieved through the estimation of asset life and the effectiveness of maintenance actions [50]. The basis besides some techniques is to split the system approach into repaired and unrepaired components when modelling the equipment reliability. These models must have into account the deterioration rate which will vary after maintenance operations such as repairs

or re-lubrication [50]. It is common for the industrial equipments to work under non-stationary conditions, with both loads and speeds changing with time or with isolated events. An effective prognostics system should allow the separations of changes in the condition caused by the deterioration of a component or due to external influences. For example, when a rotating machine increases its operating speed, the vibration level will in some cases increase which might give the perception of degradation. To surpass this problem, some authors recommend measures at the same speed of the reference, to detect faults, however this may not be possible in some applications.

As was detailed before, parameter trending is one easy and cheap way of assessing the equipment condition. However, more research needs to be conducted in order to make possible automated systems which have accurate alarm levels capable of provide reliable information. Davies [4] state that sensors with built-in hardware, capable of online processing could be a possibility.

7.3 Publications

The presented study allowed the publication of two conference papers. The first one, with the title *"Condition Monitoring of bearings under medium and low rotational speed"* was submitted to the 8th European Workshop On Structural Healthy Monitoring conference realized in Bilbao, Spain. The second one, with the title *"Comparison of vibration and acoustic measurements for detection of bearing defects"* was submitted to the Internacional Conference on Noise and Vibration Engineering, that will take place at Leuven, Belgium. The first paper can be found in appendix C, however, at the time of this thesis conclusion, the second conference was not yet realized. Therefore, to consult the published paper, the reader can visit the conference website after it is finished.

References

[1] R. B. Randall, *Vibration-based Condition Monitoring.* United Kingdom: John Wiley & Sons, Ltd, 2011.

[2] H. D. M. de Azevedo, A. M. Arajo, and N. Bouchonneau, "A review of wind turbine bearing condition monitoring: State of the art and challenges," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 368–379, 2016.

[3] A. Rai and S. H. Upadhyay, "A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings," *Tribology International*, vol. 96, pp. 289–306, 2016.

[4] A. Davies, *Handbook of Condition Monitoring Techniques and Methodology.* United Kingdom: Chapman & Hall, 1998.

[5] I. El-Thalji and E. Jantunen, "A summary of fault modelling and predictive health monitoring of rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 6061, pp. 252–272, 2015.

[6] R. Zimroz, W. Bartelmus, T. Barszcz, and J. Urbanek, "Diagnostics of bearings in presence of strong operating conditions non-stationaritya procedure of load-dependent features processing with application to wind turbine bearings," *Mechanical Systems and Signal Processing*, vol. 46, no. 1, pp. 16–27, 2014.

[7] A. R. Mohanty, *Machinery Condition Monitoring.* USA, Florida: Taylor & Francis Group, LLC, 2015.

[8] A. Bilosova and J. Bilos, *Vibration Diagnostics.* Czech Republic: Technical University of Ostrava, 2012.

[9] E. C. C. Lau and H. W. Ngan, "Detection of motor bearing outer raceway defect by wavelet packet transformed motor current signature analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 10, pp. 2683–2690, 2010.

[10] R. B. Randall and J. Antoni, "Rolling element bearing diagnosticsa tutorial," *Mechanical Systems and Signal Processing*, vol. 25, no. 2, pp. 485–520, 2011.

[11] C. Scheffer and P. Girdhar, *Machinery Vibration Analysis & Predictive Maintenance.* United Kingdom: Elsevier, 2004.

[12] J. D. Rodrigues, *Sistema Veio-Disco Modelo de Laval-Jeffcott.* Porto: FEUP, 2009.

[13] S. H. Mneney, *An Introduction to Digital Signal Processing: A Focus on Implementation.* Denmark: River Publishers ApS, 2008.

[14] W. A. Gardner, A. Napolitano, and L. Paura, "Cyclostationarity: Half a century of research," *Signal Processing*, vol. 86, no. 4, pp. 639–697, 2006.

[15] W. Cioch, O. Knapik, and J. Lekow, "Finding a frequency signature for a cyclostationary signal with applications to wheel bearing diagnostics," *Mechanical Systems and Signal Processing*, vol. 38, no. 1, pp. 55–64, 2013.

[16] N. Tandon and A. Choudhury, "A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings," *Tribology International*, vol. 32, no. 8, pp. 469–480, 1999.

[17] Z. Stamboliska, E. Rusinski, and P. Moczko, *Proactive Condition Monitoring of Low-Speed Machines*. Switzerland: Springer, 2015.

[18] J. Halme and P. Andersson, "Rolling contact fatigue and wear fundamentals for rolling bearing diagnostics - state of the art," *Engineering Tribology*, 2009.

[19] A. Gabelli, J. Lai, T. Lund, K. Rydn, I. Strandell, and G. E. Morales-Espejel, "The fatigue limit of bearing steels part ii: Characterization for life rating standards," *International Journal of Fatigue*, vol. 38, pp. 169–180, 2012.

[20] B. P. Bearings, "Bearing failure: Causes and cures," 2016.

[21] J. L. Taylor, *The Vibration Analysis Handbook: A Practical Guide for Solving Rotating Machinery Problems*, vol. US, Florida. Vibration Consultants Inc, 2003.

[22] I. O. f. Standardization, "Condition monitoring and diagnostics of machines - vibration condition monitoring," 2015.

[23] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical Systems and Signal Processing*, vol. 20, no. 7, pp. 1483–1510, 2006.

[24] I. Howard, "A review of rolling element bearing vibration "detection, diagnosis and prognosis"," *Aeronautical and Marine Research Laboratory*, 1994.

[25] M. L. Wymore, J. E. Van Dam, H. Ceylan, and D. Qiao, "A survey of health monitoring systems for wind turbines," *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 976–990, 2015.

[26] E. W. E. Association, "Wind in power - 2015 european statistics," report, 2016.

[27] F. P. Garca Mrquez, A. M. Tobias, J. M. Pinar Prez, and M. Papaelias, "Condition monitoring of wind turbines: Techniques and methods," *Renewable Energy*, vol. 46, pp. 169–178, 2012.

[28] W. Y. Liu, B. P. Tang, J. G. Han, X. N. Lu, N. N. Hu, and Z. Z. He, "The structure healthy condition monitoring and fault diagnosis methods in wind turbines: A review," *Renewable and Sustainable Energy Reviews*, vol. 44, pp. 466–472, 2015.

[29] M. Nie and L. Wang, "Review of condition monitoring and fault diagnosis technologies for wind turbine gearbox," *Procedia CIRP*, vol. 11, pp. 287–290, 2013.

[30] R. K. Mobley, *An Introduction to Predictive Maintenance*. USA: Elsevier, 2002.

[31] Z. Xia, S. Xia, L. Wan, and S. Cai, "Spectral regression based fault feature extraction for bearing accelerometer sensor signals," *Sensors Basel Switzerland*, vol. 12, no. 10, pp. 13694–719, 2012. 1424-8220 Xia, Zhanguo Xia, Shixiong Wan, Ling Cai, Shiyu Journal Article Research Support, Non-U.S. Gov't Switzerland Sensors (Basel). 2012 Oct 12;12(10):13694-719. doi: 10.3390/s121013694.

[32] A. V. Oppenheum, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. USA, New Jersey: Prentice-Hall, 1999.

[33] L. D. Enochson and R. K. Otnes, *Programming and analysis for digital time series data*. USA, California: Shock and Vibration Information Center, 1969.

[34] R. B. Blackman and J. W. Tukey, "The measurement of power spectra from the point of view of communications engineering part i," *Bell System Technical Journal*, vol. 37, no. 1, pp. 185–282, 1958.

[35] G. Heinzel, A. Rudiger, and R. Schilling, "Spectrum and spectral density estimation by the discrete fourier transfrom, including a comprehensive list of windo functions and some new flat-top windows," *Albert Einstein Institut*, 2002.

[36] V. K. Ingle and J. G. Proakis, *Digital Signal Processing Using MATLAB*. USA, Stamford: Global Engineering, 2012.

[37] M. Feldman, *Hilbert Transform Applications in Mechanical Vibration*. United Kingdom: John Wiley & Sons, Ltd., 2011.

[38] F. R. Kschischang, "The hilbert transfrom," *University of Toronto*, 2006.

[39] L. Debnath, *Wavelet Transforms and Their Applications*. USA, New York: Springer, 2002.

[40] J. L. Ferrando Chacon, V. Kappatos, W. Balachandran, and T.-H. Gan, "A novel approach for incipient defect detection in rolling bearings using acoustic emission technique," *Applied Acoustics*, vol. 89, pp. 88–100, 2015.

[41] G. L. McDonald, Q. Zhao, and M. J. Zuo, "Maximum correlated kurtosis deconvolution and application on gear tooth chip fault detection," *Mechanical Systems and Signal Processing*, vol. 33, pp. 237–255, 2012.

[42] G. Dalpiaz, R. Rubini, G. D'Elia, M. Cocconcelli, F. Chaari, R. Zimroz, W. Bartelmus, and M. Haddar, *Advances in condition Monitoring of Machinery in Non-Stationary Operations*. Germany: Springer, 2014.

[43] J. Antoni, "The spectral kurtosis: a useful tool for characterising non-stationary signals," *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 282–307, 2006.

[44] J. Antoni and R. B. Randall, "The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines," *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 308–331, 2006.

[45] A. Lesne, "Shannon entropy: a rigorous mathematical notion at the crossroads between probability, information theory, dynamical systems and statistical physics," *Mathematical Structures in Computer Science*, 2011.

[46] R. A. Fisher, *Statistical Methods for Research Workers*. United Kingdom: Oliver and Boyd, 1958.

[47] L. I. Smith, "A tutorial on principal components analysis," 2002.

[48] A. Dolgui, J. Sasiadek, M. Zaremba, S. A. Aye, P. S. Heyns, and C. J. H. Thiart, "Fault detection of slow speed bearings using an integrated approach," *International Federation of Automatic Control*, vol. 48, no. 3, pp. 1779–1784, 2015.

[49] H. Bendjama and M. S. Boucherit, "Wavelets and principal component analysis method for vibration monitoring of rotating machinery," *Journal of Theoretical and Applied Mechanics*, 2016.

[50] A. Heng, S. Zhang, A. C. C. Tan, and J. Mathew, "Rotating machinery prognostics: State of the art, challenges and opportunities," *Mechanical Systems and Signal Processing*, vol. 23, no. 3, pp. 724–739, 2009.

Appendix A

Fast Fourier Transform

The FFT is based on the number of data points N, being a power of 2. It factorizes a modified version of the matrix W_{kfn} into $\log_2 N$ matrices. The number of mathematical operations is thus reduced, as stated before, by a factor of more than 100 for a typical case where $N = 1024$ [1]. Lets introduce a modified version of the matrix W_{kfn} which will be designated matrix B, and to simplify the nomenclature, the original will be called matrix A. The new matrix B has the rows arranged in bit-reversed order as illustrated in the equation (A.1), meaning that the most significant bit is indexed, rather than the least significant one what and consequently increments the phase.

$$
\begin{array}{c}
\textit{Row}B \\
0\,0\,0\,(0) \\
0\,0\,1\,(1) \\
0\,1\,0\,(2) \\
0\,1\,1\,(3) \\
1\,0\,0\,(4) \\
1\,0\,1\,(5) \\
1\,1\,0\,(6) \\
1\,1\,1\,(7)
\end{array}
\begin{bmatrix}
\uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\
\uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \downarrow \\
\uparrow & \rightarrow & \downarrow & \leftarrow & \uparrow & \rightarrow & \downarrow & \leftarrow \\
\uparrow & \leftarrow & \downarrow & \rightarrow & \uparrow & \leftarrow & \downarrow & \rightarrow \\
\uparrow & \nearrow & \rightarrow & \searrow & \downarrow & \swarrow & \leftarrow & \nwarrow \\
\uparrow & \swarrow & \rightarrow & \nwarrow & \downarrow & \nwarrow & \leftarrow & \searrow \\
\uparrow & \searrow & \leftarrow & \nearrow & \downarrow & \nwarrow & \rightarrow & \swarrow \\
\uparrow & \nwarrow & \leftarrow & \swarrow & \downarrow & \searrow & \rightarrow & \nearrow
\end{bmatrix}
\begin{array}{c}
\textit{Row}A \\
0\,0\,0\,(0) \\
1\,0\,0\,(4) \\
0\,1\,0\,(2) \\
1\,1\,0\,(6) \\
0\,0\,1\,(1) \\
1\,0\,1\,(5) \\
0\,1\,1\,(3) \\
1\,1\,1\,(7)
\end{array}
\tag{A.1}
$$

Multiplication by B means that the results are also in bit-reversed order, but reorganize them to the correct place is a simple operation that can be done simultaneously and that performs faster than the multiplications saved with this method. The equation (A.3) shows an example for a matrix B, with $N = 8$, on the factorization into three matrices X,Y and Z. For each one of those matrices there are only two non-zero elements for each row one of each is unity [1].

$$
\begin{bmatrix}
I & I \\
I & -I
\end{bmatrix}
\tag{A.2}
$$

The factor matrices contain progressively finer rotations and the top left sub-matrix is always of the form given by equation (A.2). The factorization in powers different than 2 is also possible since the properties of the FFT are the same as the DFT.

$$
\underset{X}{\begin{bmatrix}
\uparrow & \uparrow & 0 & 0 & 0 & 0 & 0 & 0 \\
\uparrow & \downarrow & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \uparrow & \rightarrow & 0 & 0 & 0 & 0 \\
0 & 0 & \uparrow & \leftarrow & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \uparrow & \nearrow & 0 & 0 \\
0 & 0 & 0 & 0 & \uparrow & \swarrow & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \uparrow & \searrow \\
0 & 0 & 0 & 0 & 0 & 0 & \uparrow & \nwarrow
\end{bmatrix}}
\underset{Y}{\begin{bmatrix}
\uparrow & 0 & \uparrow & 0 & 0 & 0 & 0 & 0 \\
0 & \uparrow & 0 & \uparrow & 0 & 0 & 0 & 0 \\
\uparrow & 0 & \downarrow & 0 & 0 & 0 & 0 & 0 \\
0 & \uparrow & 0 & \downarrow & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \uparrow & 0 & \rightarrow & 0 \\
0 & 0 & 0 & 0 & 0 & \uparrow & 0 & \rightarrow \\
0 & 0 & 0 & 0 & \uparrow & 0 & \leftarrow & 0 \\
0 & 0 & 0 & 0 & 0 & \uparrow & 0 & \leftarrow
\end{bmatrix}}
\underset{Z}{\begin{bmatrix}
\uparrow & 0 & 0 & 0 & \uparrow & 0 & 0 & 0 \\
0 & \uparrow & 0 & 0 & 0 & \uparrow & 0 & 0 \\
0 & 0 & \uparrow & 0 & 0 & 0 & \uparrow & 0 \\
0 & 0 & 0 & \uparrow & 0 & 0 & 0 & \uparrow \\
\uparrow & 0 & 0 & 0 & \downarrow & 0 & 0 & 0 \\
0 & \uparrow & 0 & 0 & 0 & \downarrow & 0 & 0 \\
0 & 0 & \uparrow & 0 & 0 & 0 & \downarrow & 0 \\
0 & 0 & 0 & \uparrow & 0 & 0 & 0 & \downarrow
\end{bmatrix}}
$$

$$
= \underset{B}{\begin{bmatrix}
\uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\
\uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \downarrow \\
\uparrow & \rightarrow & \downarrow & \leftarrow & \uparrow & \rightarrow & \downarrow & \leftarrow \\
\uparrow & \leftarrow & \downarrow & \rightarrow & \uparrow & \leftarrow & \downarrow & \rightarrow \\
\uparrow & \nearrow & \rightarrow & \searrow & \downarrow & \swarrow & \leftarrow & \nwarrow \\
\uparrow & \swarrow & \rightarrow & \nwarrow & \downarrow & \nwarrow & \leftarrow & \searrow \\
\uparrow & \searrow & \leftarrow & \nearrow & \downarrow & \nwarrow & \rightarrow & \swarrow \\
\uparrow & \nwarrow & \leftarrow & \swarrow & \downarrow & \searrow & \rightarrow & \nearrow
\end{bmatrix}} \tag{A.3}
$$

Appendix B

Mounting Procedures

B.1 FMTC - Low Speed Test Bench

The procedure which should be followed, when measuring in the FMTC facilities and to ensure the best repeatability as well as accuracy of the measure quantities, is the one described below.

1. Remove the top cover of the gearbox by unscrewing the 4 bolts at the corners of the same.
2. Unscrew the bolts that fix the bearings to the shaft of interest, in this case, the middle one, on both ends to allow the axial movement of the shaft.
3. Totally remove the shaft encoder placed on the input shaft and place it away in order to clear the area for the bearing housing removal.
4. Locate the screws that fix the bearing housing marked in figure B.1 and unscrew them completely.

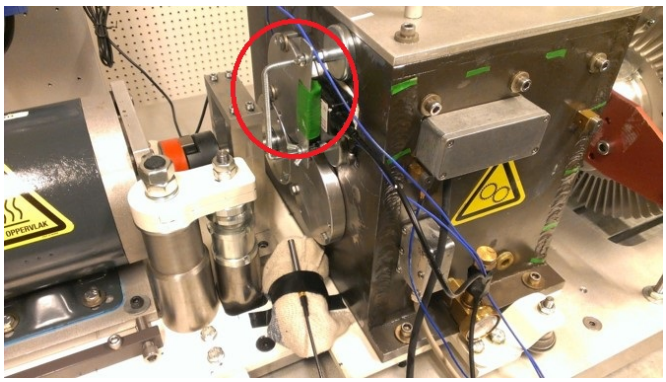


Figure B.1: FMTC Procedure - Bearing housing location

5. In the bearing housing attach the removed screws in the holes that were previously empty. Fasten the screws until those touch the surface of the gearbox. With small increments use the screws to dismount the bearing housing by making it slid off its position.
6. Inspect if the mounted extremity of the shaft is not forcing and consequently damage the respectively bearing housing. If needed take actions to assure that the shaft will not damage the structure or itself.

7. With the bearing housing in a flat surface remove the previous bearing and the spring. Clean all the surfaces with an appropriate cleaning solution and check for wear particles.

8. Place the spring back into its place. Check the condition of the faulty bearing to mount and mount it in the housing with a flat object and a hammer. The bearing has to be mounted with the minimum misalignment between the central axis of the housing and the bearing itself. An error in this step could lead to surface damages and improper load distribution. If the case in study is the outer race fault bearing, place the fault directly under the axis to maximize the load on the fault.

9. Replace the bearing housing into the gearbox through the assembly of the shaft in the bearing. Use your hands to push the housing as much as possible.

10. Fix the centring screw in the top of the housing and attach it slightly. After this, also slightly, attach the remaining two bolts.

11. In small increments, fasten all the screws into a clockwise direction until the bearing housing is in its final position.

12. Verify the alignment between the gears marked in figure B.2. If not properly aligned make the corrections needed. This step minimize the gearmesh frequency components amplitudes.

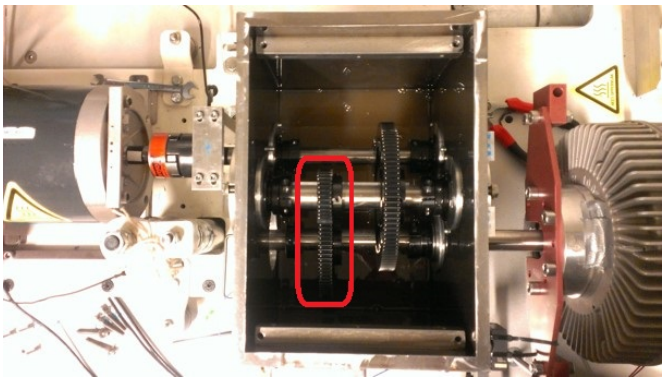


Figure B.2: FMTC Procedure - Gears alignment

13. Firmly fasten the bearings to the shaft as symmetrically as possible.

14. Check the oil level, correct if necessary and close the gearbox.

B.2 SpectraQuest - Medium Speed Test Bench

The procedure which should be followed, when measuring in the SQ test bench and to ensure the best repeatability as well as accuracy of the measure quantities, is the one described below.

1. After stop the driver, wait 10 seconds. Loose the bolts that secure the bearings to the shaft as well as the ones in the load mass if necessary.

2. Move the shaft axially to uncouple the motor from the shaft.

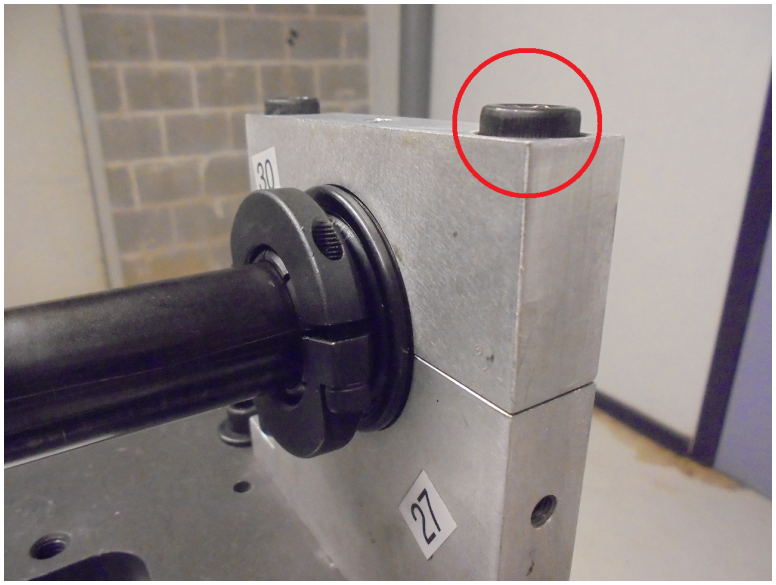


Figure B.3: SQ Procedure - Bolts location

3. Remove the superior part of the bearing housings by loosing the bolts marked in figure B.3.

4. Carefully remove the shaft from the structure and place it in a steady platform. Remove the previous bearing and place the new one having into account the fault location in the case of the outer race fault.

5. Replace the shaft in the structure with the coupling open and close the bearing housings with a slight fasten to the screws.

6. Move the shaft against the motor until near its final mounted position.

7. With a dinamometric tool as shown in figure B.4 fasten the housings screws in a cross direction with a torque of 22 Nm.

8. Move the shaft against the motor to its final position making sure that it is not forcing the motor which would cause internal bearing damage.

9. Firmly secure the bearings and the load to the shaft achieving as much symmetry as possible.

10. Release all the 4 bolts, marked with the read circle in figure B.5, that secure the plate, where the shaft is mounted, to the test rig base.

11. Starting with the two calibration screws, marked with the yellow circle in figure B.5, which are located in the axial direction of the shaft, move the structure in a way that minimizes gaps in the coupling without forcing the motor shaft. Slightly secure the calibration screws.

12. With the transversal calibration screws which are located in the opposite end of the motor, minimize the angular misalignment in the structure. After that, use the transversal calibration screws located close to the engine to refine the structure position. This process is iterative and will have a large influence in the final results. After an acceptable compromise is achieved slightly secure the calibration screws.

Figure B.4: SQ Procedure - Fasten example

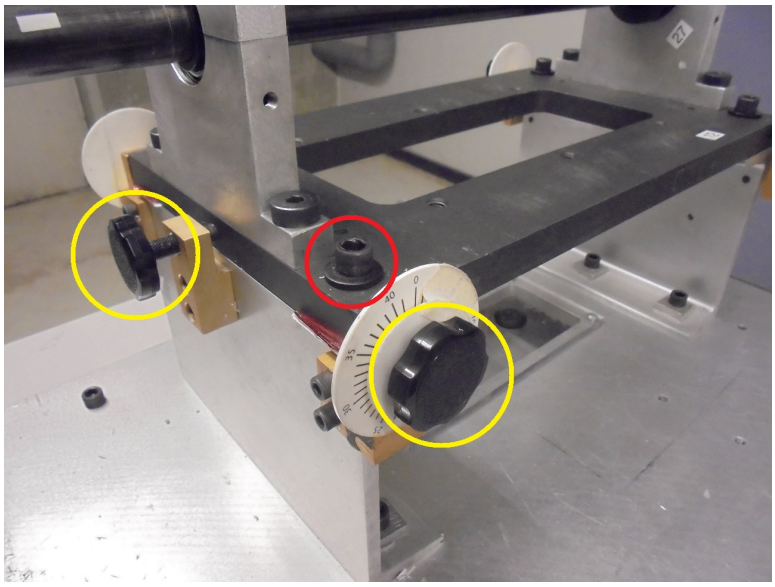


Figure B.5: SQ Procedure - Base plate and calibration screws

13. Starting with the axial calibration screws, moving to the ones further from the motor and finally to the ones close to the motor, firmly fasten all the calibration screws to lock the structure position.

14. Fasten the bolts marked with the red circle in figure B.5, using a dynamometric tool with a torque of 10 Nm.

Appendix C

Research Paper - Condition monitoring of bearings under medium and low rotational speed

Condition monitoring of bearings under medium and low rotational speed

Carina Freitas [1], Paulo Morais [1], Jacques Cuenca [1], Agusmian Partogi Ompusunggu [2], Mathieu Sarrazin [1], Karl Janssens [1]

[1] Siemens Industry Software NV
Leuven 3001 BELGIUM carina.freitas@siemens.com, paulojorge.morais@hotmail.com, jacques.cuenca@siemens.com, mathieu.sarrazin@siemens.com, karl.janssens@siemens.com
[2] Flanders Make vzw
Leuven 3001 BELGIUM agusmian.ompusunggu@flandersmake.be

Abstract
Rolling bearings are an important component in rotating machinery and their failure is one of the main causes of machine breakdown. It has been proven that diagnosis methods based on envelope analysis of vibration signals enable the detection of bearing faults in medium to high speed regimes. However, it has been observed that these methods do not provide a satisfactory performance for diagnosing bearings at low rotational speed. One of the known reasons for this is the usage of inadequate monitoring techniques and measuring instruments that are not sensitive enough to low frequency signature. A major challenge for fault diagnosis at low rotational speeds is their slow evolution in time and the weak amplitude of the induced-fault vibrations. This paper aims at presenting the techniques proposed by Randall and co-workers that allow semi-automated processing of machine vibration signals to diagnose bearing faults under different operational conditions for low and medium rotational speed. This work includes a comparative analysis of the benefits of applying the different methods for obtaining statistical features in the time domain. The techniques under interest are the cepstral editing method for the removal of unwanted components masking the fault signature, and minimum entropy deconvolution, which allows recovering an estimation of the original impulsive signal originating from the fault by removing the transfer path between the fault and the sensor location. In order to understand the limitations of the methods, measurements under different speeds and loads were performed. The goal is to find key features that allow the detection of bearing faults such as outer race and inner race defects at different speed regimes. The fault diagnosis and prognosis of low speed machines remain a topic of interest since vibration analysis generally fails to detect vibrations arising from faulty components. A further complicating aspect is the fact that machines that work in this regime are generally characterized by varying rotational speed and load, thus the diagnostic features to extract must be load and speed dependent.

Keywords: Vibration Analysis, Bearings, Gearbox, Condition Monitoring, Medium and Low Rotational Speed

1. INTRODUCTION

Bearings are widely used in industry and their deterioration is one of the most common causes of failure in rotating machinery. Recent studies have aimed at identifying and applying techniques enhancing and detecting bearing faults. The most well-established monitoring technique is vibration analysis for medium to high speed regimes [1] [2]. Nevertheless, it has been reported that the identification of bearing faults through vibration signals under low rotational speed is more challenging due to the low change in signal energy

generated from the faults and may not be detectable using traditional accelerometers. To overcome this issue an alternative technology based on acoustic emission sensing has been used providing more satisfactory results than classical vibration analysis for this speed range [3] [4] [5].

In many industries such as steel and paper mills, biological applications and wind turbines, it is common to find machines operating at low speed regimes. To increase the complexity, this type of machines is normally associated to varying operational conditions [3] [5].

One of the main domains of research on bearing faults is the study of wind turbines, which operate in a wide range of speeds. It is important to understand that the current commercial systems remain very expensive and issues with prediction reliability and accuracy persist [2]. To overcome some of these problems, an alternative approach was proposed in Ref. [2], by using indicators which are not sensitive to these operational conditions but to the change in health of the gearbox. In such study the authors defended the importance of pre-processing the signals before performing the feature extraction. The goal of the pre-processing is to filter out noise and normalise operational variability. It has been proven that root mean square (RMS) and peak values are good indicators for gearbox health monitoring. However, it was pointed out that RMS will be only suitable for intermediate/high speed.

Several studies have been performed in order to identify the features that better identify a bearing fault. In [6], the author studied statistical parameters such as root mean square, crest factor, skewness and kurtosis for a test rig operating from 500 to 3500 rpm. The study led to identifying certain trends per rotational speed range for kurtosis, and kurtosis versus crest factor. Four different bearing conditions (healthy, inner and outer race fault and ball defect) were analysed. This author concluded that the statistical parameters were rotational speed dependent. In [7] a study was performed to verify the variations of statistical moments to identify damage at a much earlier stage. An important step introduced in this analysis was the removal of unwanted noise using digital filtering. Four bearing condition (healthy, inner and outer race fault and ball defect) were analysed by means of kurtosis and skewness and five test speeds between 1000 and 3000 rpm were selected.

The authors defended in [1] [8] the need of applying some signal pre-processing in order to enhance the bearing fault signature in the vibration signals. A vibration signal collected from the bearing housing is composed of different components that might mask the bearing fault signature. The authors confirmed that this procedure was successfully applied in the majority of the cases and it could be as well applicable for low speed. The first step is the removal of unwanted deterministic components, and then it is important to recover the impulsiveness of the signal at the source by removing the transfer path between the sensor and the location where the fault occurred. After this pre-processing, the final step is the application of the traditional envelope analysis.

The goal of this work is to combine the procedure developed in [8] with the application of statistical features [2] [6] [7] in the time domain for inner and outer race faults in medium speed regime. A validation of the procedure to detect inner race fault under low rotational speed is also included.

2. METHOD

2.1 Vibration Signature of Bearing Faults

Vibration signals generated by bearing faults are impulsive, at least at the source. However, the vibration signal measured at the bearing housing is the result of a combination of different types of signals determined by gears, transmission path and measurement noise.

Shaft and gear related signals are deterministic and dominant, thus representing one of the most important masking components. The other source that needs to be taken into account is the transmission path from the source to the sensor which can in turn yield a decrease in the impulsivity of the signal.
This paper intends to study the benefits of different signal pre-processing techniques, namely cepstrum editing and minimum entropy deconvolution (MED) for bearing fault identification in different speed and load regimes (Figure 1).

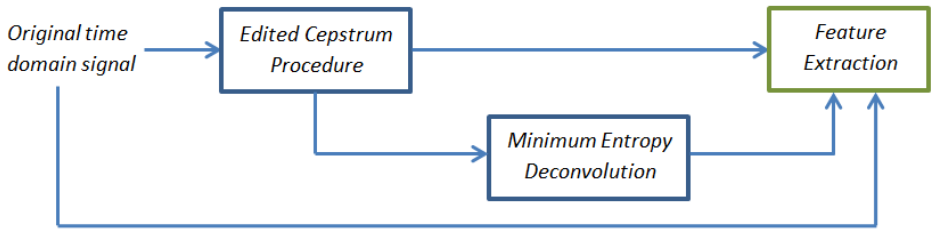


Figure 1: Schematic representation of the procedures under analysis

2.2 Deterministic Components Removal

Different techniques are available in literature, which allow the removal of the deterministic components (linear prediction, adaptive noise cancellation, self-adaptive noise cancellation, discrete/random separation, time synchronous averaging, cepstral editing procedure) [1] [9] [10]. The investigated method in this paper is the cepstral editing procedure (CEP) due to the efficiency and ease of interpretation and implementation. It has been shown in literature that CEP enhances the bearing fault related signals more significantly than other techniques [9] [10]. It removes selected discrete frequency components, including sidebands.
The most common definition of cepstrum is the *inverse Fourier Transform of the logarithmic spectrum* of a signal ***f(t)***, which can be written as

$$C(\tau) = \Im^{-1}\{ln(F(f))\}, \quad (1)$$

$$F(f) = \Im\{f(t)\} = A(f)e^{j\varphi(f)}, \quad (2)$$

where ***F(f)*** is the frequency spectrum of ***f(t)***, $\Im^{-1}$ the inverse Fourier Transform and τ is the quefrency.
If the phase is retained, the complex cepstrum, C_c, has logarithmic amplitude as real part and phase as imaginary part. If the phase is disregarded, the so-called 'real cepstrum' $\boldsymbol{C}_r$ is obtained:

$$C_c(\tau) = \Im^{-1}\{ln(A(f)) + j\varphi(f)\}, \quad (3)$$

$$C_r(\tau) = \Im^{-1}\{ln(A(f))\}, \quad (4)$$

Figure 2 shows the procedure used to remove the deterministic components. The main advantage of cepstral editing for the present application is that the unwanted deterministic components are removed in a subtractive manner.

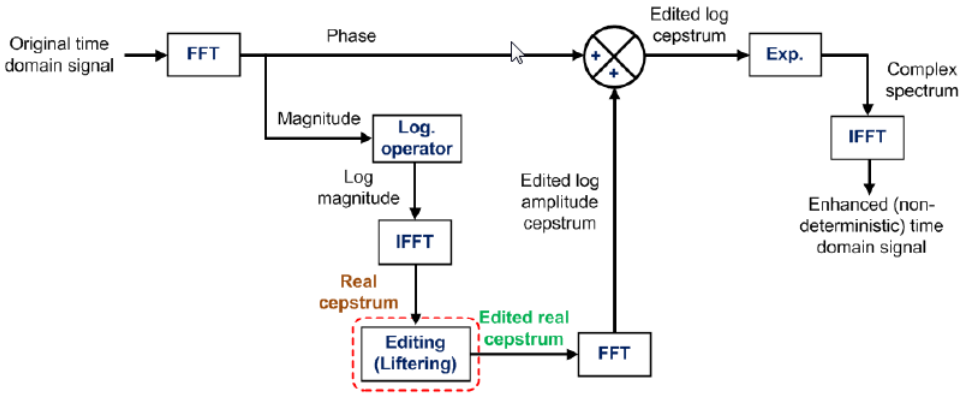


Figure 2: Schematic representation of the cepstral method to remove deterministic components [9]

2.3 Transmission Path Removal

Minimum entropy deconvolution (MED) is an existing method allowing the removal of the effect of the transmission path. This technique is a type of adaptative filtering that searches for an optimum set of filter coefficients that recover the output signal (or inverse filter) with the maximum value of kurtosis [11].

As referred by Randall [1], this method permits to reconstruct the shape of the original impulses at the source position. By minimizing the entropy, the kurtosis of the inverse filter output is maximized. Figure 3 illustrates the concept behind the MED. The forcing signal *e(n)* passes through the structural filter **h** whose output is mixed with noise *u(n)* and gives the measured output *x(n)*. The inverse filter **f** produces the output *y(n)* which has to be as close to the original input *e(n)* which is unknown but should be as impulsive as possible.

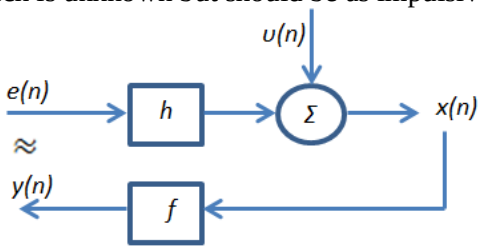


Figure 3: Inverse filtering (deconvolution) process for MED

The filter **f** is modelled as a FIR filter with L coefficients. **f** has to invert the system **h**.

$$y(n) = \sum_{l=1}^{L} \mathbf{f}(l)x(n-1) \tag{5}$$

The objective function to be maximized is the kurtosis of the output signal *y(n)* through the filter **f**. The maximum is found through the coeficients of the filter **f** by making the derivative of the following objective function equal to zero:

$$O_k(f) = \frac{\sum_{n=0}^{N-1} y^4(n)}{[\sum_{n=0}^{N-1} y^2(n)]^2} \qquad \frac{\partial O_k(f)}{\partial f} = 0 \tag{6}$$

2.4 Feature Extraction

For this paper eight statistical features [12] are considered. These are applied to the vibration signals in the time domain after pre-processing and are detailed as follows:

Root Mean Square: the square root of the arithmetic mean of the squares of the original signal.

$$RMS = \sqrt{\frac{1}{n}\sum_{k=1}^{n} x_k^2} \quad (7)$$

Variance: a measure of the spread of a signal from its mean.

$$Variance = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2 \quad (8)$$

Kurtosis: a measure of the peakedness of the vibration signal. *s* corresponds to the standard deviation.

$$Kurtosis = \frac{\sum_{k=1}^{n}(x_i - \bar{x})^4}{(n-1)s^4} \quad (9)$$

Crest Factor: the ratio of the peak value to the RMS value. It gives the shape of the waveform.

$$Crest\ Factor = \frac{Peak\ value}{RMS} \quad (10)$$

Peak to Peak: the range between the maximum and minimum value in the signal.

$$Peak\_to\_Peak = \max(x) - \min(x) \quad (11)$$

Shannon Entropy [13]: Entropy is a quantitative measure of disorder. It quantifies the average amount of information in the signal.

$$Ent_{shannon} = -\sum_{k=1}^{n} x_k^2 \log(x_k^2) \quad (12)$$

Log energy entropy [13]: log energy entropy in the signal.

$$Ent_{log} = \sum_{k=1}^{n} \log(x_k^2) \quad (13)$$

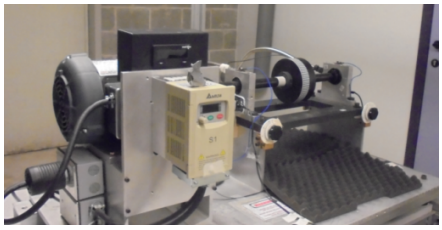
75th percentile: $x_{0.75}$ is the value below which 75% of the data falls.

$$0.75 = P(x \leq x_{0.75}) \quad (14)$$

3. EXPERIMENTAL SETUP

3.1 Machine Fault Simulator SpectraQuest

The measurements for medium speed regime were performed in a fault machine simulator from SpectraQuest (Figure 4). The machine is driven by an electric motor controlled by a simple variable frequency drive. Three different bearing conditions were measured (healthy, inner and outer race fault) for a speed range between 780 and 2340 rpm with an interval of 120 rpm. A 5 kg static load was added and located in the middle of the shaft (LM) and at 3 cm from the bearing under analysis (LR). The bearings under analysis were ER-16K.



Bearing ER-16K	
Rolling element diameter	7.94 mm
Pitch diameter	39.32 mm
Ball number	9
Contact angle	0^o

Figure 4: Fault Machine Simulator SpectraQuest and bearing characteristics

3.2 Gearbox Setup

The measurements for low rotational speed were performed in the test rig depicted in Figure 5. The setup consists of one gearbox which is driven through a flexible coupling connected to an induction motor and controlled by a variable frequency drive. The input shaft was set to run at 480, 1200 and 2400 Hz. The bearing under analysis was located on the intermediate shaft of the gearbox which has a gear ratio of 29/100. The rotational speed of this shaft was 139, 348 and 696 rpm, correspondingly. The bearings under analysis were ER-16K and were measured in two different conditions, healthy and with an inner race fault. The output shaft is coupled with a magnetic-particle break, where the torque was adjusted to 25% (T_{25}) and 100% (T_{100}) of 27.5 Nm.

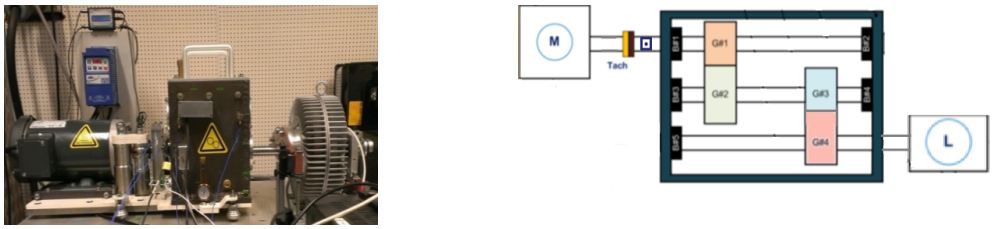


Figure 5: Picture and schematic diagram of the gearbox setup

4. DATA ANALYSIS AND DISCUSSION

4.1 Medium Rotational Speed Analysis

To understand the benefits of the different pre-processing techniques, vibration signals were acquired and analyzed for different speed and load conditions on the machine fault simulator of SpectraQuest (Figure 4). Figure 6 shows a representative case of this analysis for the kurtosis feature calculated in the healthy and inner race fault cases. Focusing on the feature value in the healthy case, as expected, the kurtosis level is much lower (~3) compared to the faulty case. Still it is possible to verify the benefits of the MED for this feature, due to a higher difference in the values when compared with the reference (healthy). It is possible as well to verify that the two loads present comparable trends. Moreover, it is relevant to add that for this processing the tuning parameters for the cepstrum were a liftering band of 15% and for the MED a filter length of 200 samples. The stop criterion was chosen as a difference in kurtosis of 0.1 between successive iterations. The same conclusions are valid for the outer race fault, with a better trending with the load located in the middle of the shaft. The benefits of the MED were also noticed for the crest factor for both faults.

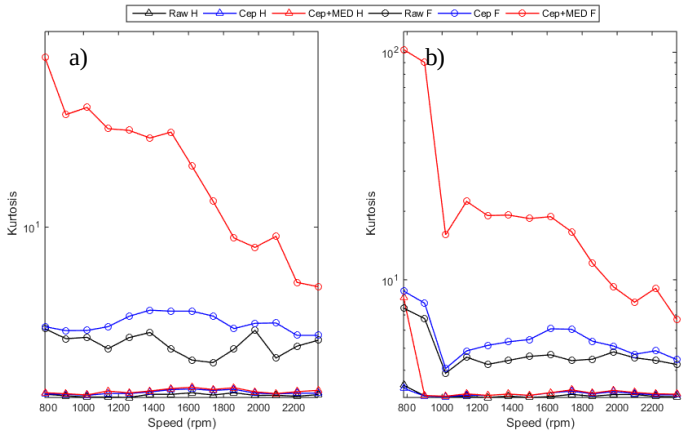


Figure 6: Kurtosis for Inner Race Fault vs. healthy case: a) Load Middle, b) Load Right. Legend for figures 6 to 8 and 11 to 13: , without pre-processing for healthy case; , with application of edited cepstrum for healthy case; , with application of edited cepstrum and MED methods for healthy case; , without pre-processing for the faulty case; with application of edited cepstrum and MED methods for faulty case; with application of edited cepstrum and MED methods for faulty case

Figure 7 presents the calculated RMS feature for the inner race fault versus the healthy case for two different load conditions. In this case and for the other statistical features (Variance, Peak to Peak, Percentile 75 and Entropy) it is concluded that there is no benefit of applying the MED method. There is already a clear distinction between the faulty and healthy feature value without any signal pre-processing. For this case, the application of the cepstrum editing method did not provide any other extra advantage in further increasing the deviation of faulty from healthy results. Figure 8 shows the RMS values for the outer race fault which confirm the previous observations. For the test rig shown in Figure 4, the load level does not seem to influence the trends but it is not possible to draw a definitive conclusion about the influence of the load since it is light (5 kg) and does not represent the loads in the industrial environment. In a "real-life" application the influence of the load might be more significant. The other important aspect to be noted from this investigation is the fact that the feature values are speed-dependent and the method of the static threshold will not provide satisfactory results and might lead to false alarms. Therefore, the threshold should be set as a function of the rotational speed.

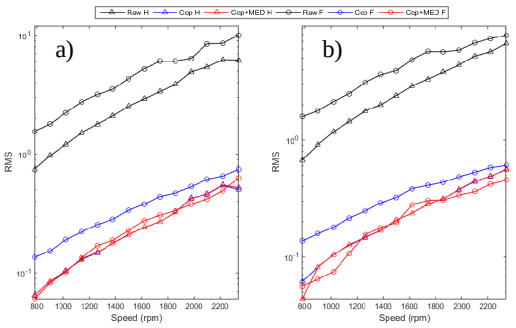


Figure 7: RMS for Inner Race Fault vs. healthy case: a) Load Middle, b) Load Right. Same legend as Figure 6

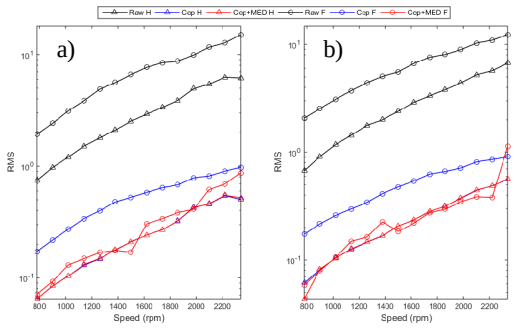


Figure 8: RMS for Outer Race Fault vs. healthy case: a) Load Middle, b) Load Right. Same legend as Figure 6

Figure 9 provides the correlation matrix of the considered features. The upper triangle corresponds to the Load Right (LR) and the lower triangle to the Load Middle (LM) for an

inner race fault case. Figure 10 shows the results for the outer race fault case. There are general conclusions that can be taken independently of the type of fault and load: the parameters Variance, RMS, Percentile 75 and Shannon Entropy provide the higher correlations. With the application of the cepstral editing method and the MED, it can be observed that the number of highly correlated variables is reduced to three, leaving out the $Ent_{shannon}$. The use of this processing reduces the correlation coefficient between variables.

An aspect of interest for fault detection is the identification of patterns by using a minimum set of uncorrelated features. Based on the present results, it is worth noting that a set of correlated features provide redundant information for fault detection purposes. Thus, in that case the consideration of one feature should be enough to evaluate the trend with speed. A particular case was observed for the raw signal of inner race fault with the load located in the middle. For this configuration a high correlation was observed between the Peak-to-Peak feature and all parameters except for the Crest Factor and Kurtosis features. Kurtosis and Crest Factor are the variables providing the lowest correlation coefficients in comparison with the other statistical features for the presented fault simulator test rig. However, the interdependence between features may be structure-dependent.

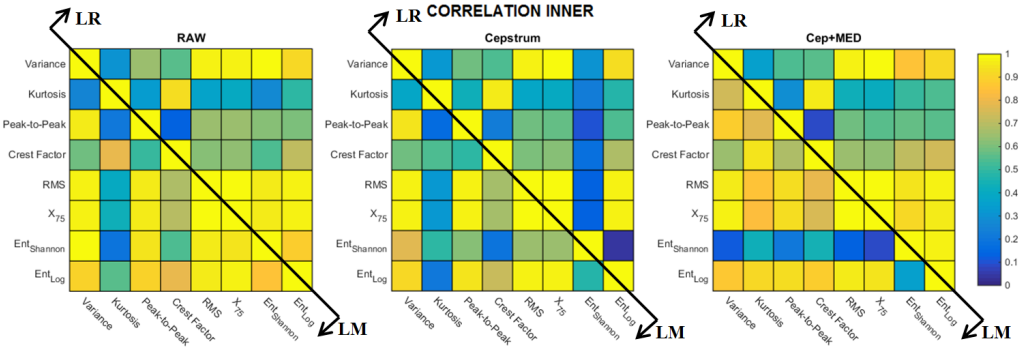


Figure 9: Correlation Coefficient of the features for Inner Race Fault for two load positions left) without processing, middle) with cepstrum editing, right) with cepstrum editing and MED

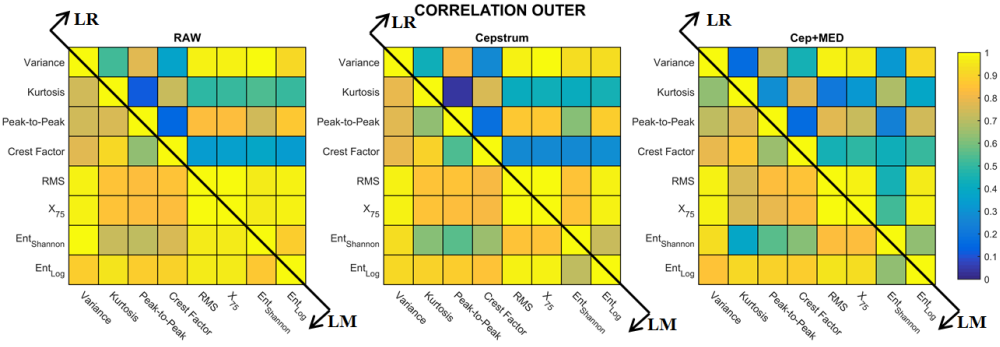


Figure 10: Correlation Coefficient of the features for Outer Race Fault for two load positions left) without processing, middle) with cepstrum editing, right) with cepstrum editing and MED

4.2 Low Rotational Speed Analysis

A low rotational speed analysis was conducted on the gearbox setup (Figure 5). For the deterministic components removal with the cepstrum editing technique, a liftering band of 15% and a MED filter length of 120 with stop criteria of 2.5 were applied in order to enhance

the impulsive components. For this specific dataset it could be noticed that the appropriate MED filter must be designed carefully in order to prevent enhancing parasite impulsive components. For instance, this could lead to high values of kurtosis for a healthy bearing which could give false alarms.

Figures 11 and 12 show a small difference in the feature values between the faulty and healthy condition without pre-processing. The difference is small at the lowest speed and tends to increase with rotational speed. In Figure 11 the benefits of using extra processing can be observed; namely, the removal of the discrete components enhances the bearing fault signature in the signal. These conclusions are valid for both load conditions. As previously concluded for the fault simulator machine and for the same reason, it is not possible to draw here any conclusion regarding the effect of the load on the feature values. Once again it is very clear from the figures that the features are speed dependent.

Figure 12 presents the RMS for both torques, where, as before, it proves difficulties to identify a fault without pre-processing for certain speeds. Figure 13 shows the feature Peak-to-Peak and for this particular case the fault can be clearly identified without processing. For this case the application of the MED may lead to incoherent results (lowest speed for T_{25}) but the cepstrum helps to enhance the fault.

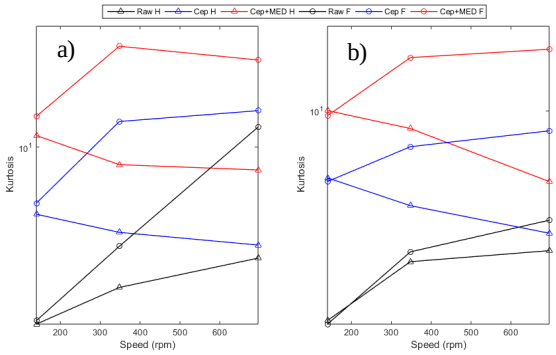


Figure 11: Kurtosis for Inner Race Fault vs. healthy case: a) $T_{25,}$ b) T_{100}. Same legend as Figure 6

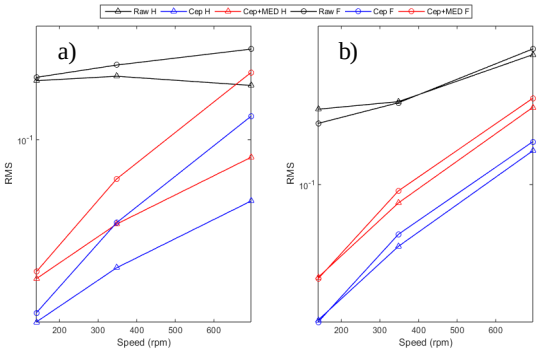


Figure 12: RMS for Inner Race Fault vs. healthy case: a) $T_{25,}$ b) T_{100}. Same legend as Figure 6

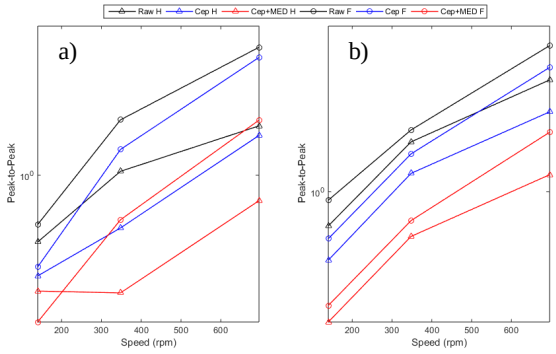


Figure 13: Peak to Peak for Inner Race Fault vs. healthy case: a) $T_{25,}$ b) T_{100}. Same legend as Figure 6

5. CONCLUSIONS

This paper was intended to study the benefits of different signal pre-processing techniques, namely cepstrum editing and minimum entropy deconvolution (MED) for bearing fault identification in different speed and load regimes. This was performed by extracting several features from vibration signals acquired on a bearing fault simulator

setup for different speed regimes and loads. For medium speed the benefit of the pre-processing techniques is not substantial. Only the kurtosis and crest factor seemed to benefit from it. The same conclusions apply for both inner and outer race faults. It was as well verified that several features were correlating with each other, and it might be important to consider the application of feature selection algorithms for pattern recognition proposes. Finally, for low rotational speed (gearbox setup) it was observed that several features (e.g. RMS, Kurtosis, Variance, Ent_{Log}) benefit from the use of minimum entropy deconvolution and especially of the cepstrum. This observation can be related or driven by the complexity of the test rig. However, it is important to validate these conclusions by performing measurements at additional speeds in order to obtain a clearer view of the trending since this analysis was limited to three speed points. As a final remark, the selection of the MED parameters needs to be done carefully in order to prevent false alarms in the fault detection. New tests need to be executed to understand if the benefits of applying pre-processing are directly related to the complexity of the test rig.

6. REFERENCES

[1] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics - A tutorial," *Mechanical Systems and Signal Processing,* vol. 25, pp. 485-520, 2011.

[2] J. Igba, K. Alemzadeh, C. Durugbo and E. T. Eiriksson, "Analysing RMS and peak values of vibration signals for condition monitoring fo wind turbine gearboxes," *Renewable Energy,* vol. 91, pp. 90-106, 2016.

[3] B. Van Hecke, J. Yoon and D. He, "Low speed bearing fault diagnosis using acoustic emission sensors," *Applied Acoustics,* vol. 105, pp. 35-44, 2016.

[4] N. Jamaludin and D. Mba, "Monitoring of extremely slow rolling element bearings: part I," *NDT & E International,* vol. 35, pp. 349-358, 2002.

[5] A. Widodo, E. Y. Kim, J.-D. Son, B.-S. Yang, A. C. C. Tan, D.-S. Gu, B.-K. Choi and J. Mathew, "Fault diagnosis of low speed bearing based on relevance vector machine and support vector machine," *Expert Systems with Applications,* vol. 36, pp. 7252-7261, 2009.

[6] R. B. W. Heng and M. J. M. Nor, "Statistical Analysis of Sound and Vibration Signals for Monitoring Rolling Element Bearing," *Applied Acoustics,* vol. 53, pp. 211-226, 1997.

[7] H. R. Martin and F. Honarvar, "Application of Statistical Moments to Bearing Failure Detection," *Applied Acoustics,* vol. 44, no. 1, pp. 67-77, 1995.

[8] N. Sawalhi and R. B. Randall, "Semi-Automated bearing diagnostics - Three case studies," in *International Congress on Condition Monitoring and Diagnostic Engineering Management*, Faro, 2007.

[9] A. P. Ompusunggu and T. A. Bartic, "Automated cepstral editing procedure (ACEP) for removing discrete components from vibration signals," in *International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, Oxford, 2015.

[10] R. B. Randall, N. Sawalhi and M. Coats, "A comparison of methods for separation of deterministic and random signals," *The International Journal of Condition Monitoring,* vol. 1, no. 1, pp. 11-19, 2011.

[11] N. Sawalhi, R. B. Randall and H. Endo, "The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis," *Mechanical Systems and Signal Processing,* vol. 21, pp. 2616-2633, 2007.

[12] A. D. Poularikas, Handbook of Formulas and Tables for Signal Processing, CRC Press LLC, 1998.

[13] R. R. Coifman and M. V. Wickerhauser, "Entropy-based Algorithms for best basis selection," *IEEE Trans. on Inf. Theory,* vol. 38, no. 2, pp. 713-718, 1992.